Potenz-Gesetze im Web – auch im @grarbereich

MICHAEL CLASEN, KIEL
ROLF A.E. MÜLLER, KIEL

Abstract
The web has grown into a seemingly chaotic tangle of websites, all connected by hyperlinks. Although the complexity of the web defies any detailed description, methods borrowed from statistical mechanics have proved to be useful for analyzing the web. In this article we introduce into the literature dealing with power-law-phenomena and compare existing results with an own small study of German agricultural Web-Sites.

1 Einleitung

Das rapide Wachstum des World Wide Web ist legendär – kein anderes Informationsnetzwerk hat sich jemals so schnell und so weit verbreitet wie das Web. Das Ergebnis des ungeplanten und von keiner zentralen Instanz gesteuerten Wachstums ist ein unüberschaubares Universum von Galaxien von Websites. Es dauerte jedoch nicht lange, bis Wissenschaftler versuchten Ordnung inmitten des scheinbaren Chaos zu finden. Sie hatten Erfolg. Auf der Grundlage von graphentheoretischen Modellen gelang es, Methoden zur Messung von Eigenschaften des Webs zu entwickeln. Diese Messungen ergaben für das Web empirische Regelmäßigkeiten, die der naiven Wahrnehmung eines unstrukturierten und unübersichtlichen Netzwerks widersprechen. Ein wichtiges Beispiel dieser Regelmäßigkeiten ist die sogenannte Potenzverteilung einer Reihe meßbarer Phänomene im Web, wie z.B. die Zahl der eingehenden oder ausgehenden Hyperlinks einer Website, die Zahl der Seiten einer Site oder auch die Zahl der Besucher einer Website.
Die meisten dieser Studien wurden für große Ausschnitte des Webs durchgeführt. Wir haben die Hypothese der Potenzverteilung anhand von Messungen für einen vergleichsweise sehr kleinen Ausschnitt des Web angewendet: die Zahl der Besucher von landwirtschaftlichen Websites, die ihre Seitenaufrufe durch die Website Land24-Hitparade zählen lassen. Zu unserer Überraschung zeigte sich, daß auch für diesen kleinen Ausschnitt des Webs die Potenz-Gesetz-Eigenschaft zutrifft.
Unser Beitrag gliedert sich in drei Haupteile. Im nächsten Kapitel fassen wir kurz die wichtigsten Ansätze zur Messung des Webs zusammen. Danach stellen wir im dritten Abschnitt unsere Daten und Erhebungsmethode dar und berichten unsere Ergebnisse. Im vierten Abschnitt diskutieren wir die Bedeutung unserer Ergebnisse für die empirische Forschung über die Landwirtschaft im Web und für die Praxis des land- und ernährungswirtschaftlichen
E-Business.

2 Potenzgesetze beherrschen das Web

Die hohe Komplexität des scheinbar chaotisch aufgebauten Internets macht seine detaillierte Beschreibung nahezu unmöglich. Als nützlich zur Beschreibung der Struktur des Webs haben sich dagegen Methoden erwiesen, die ursprünglich aus der statistischen Mechanik stammen. Einen guten Einblick in die bisherige Forschung geben ALBERT UND BARABÁSI (2001), BRODER et al. (2000), HUBERMAN (2001), KLEINBERG UND LAWRENCE (2001) oder BARABÁSI (2001), wobei die beiden erstgenannten Arbeiten stark mathematisch orientiert sind. Dem interessierten Laien sei BARABÁSI (2002) empfohlen.
In der statistischen Webanalyse wird das Web gewöhnlich als Graph modelliert, bei dem einzelne Webseiten oder ganze Websites die Knoten und die Links zwischen den Webseiten

die Kanten bilden. In unserer Untersuchung folgten wir dem Ansatz von ADAMIC (n.d.) und definierten Websites als die Knoten des Graphen und die temporären Verbindungen des Browsers eines Nutzers zu der besuchten Website als die Kanten.
Zur Beschreibung komplexer Graphen werden üblicherweise drei Maße verwendet: a) die durchschnittliche Pfadlänge, b) der Clusteringkoeffizient und c) der Vernetzungsgrad. In unserer Untersuchung verwenden wir ausschließlich den Vernetzungsgrad k eines Knoten, der als die Anzahl der Kanten eines Knoten definiert ist. In einem Netzwerk mit n Knoten bedeutet k = 0, daß der betreffende Knoten isoliert ist und k = n-1, daß der Knoten mit allen anderen Knoten verbunden ist.
In großen Netzwerken ist die Verteilung von k von besonderem Interesse und im Fall des Vernetzungsgrads der Knoten im Web (Webseiten oder Sites) hat es sich gezeigt, daß k zumeist von einer Potenzverteilung der Form $P(k) \sim k^{-\gamma}$ beschrieben wird, wobei P(k) die Wahrscheinlichkeit darstellt, daß ein zufällig ausgewählter Knoten exakt k Kanten aufweist. In empirischen Studien wurden für das Internet γ-Werte zwischen 1,94 und 2,72 ermittelt, wobei die genauen Werte mit der Definition der Knoten (Hosts, Webseiten, Websites, etc.) und der Kanten (Weblinks, Webhits, etc.) variierten. Potenzverteilungen dieser Art sind in mehreren Untersuchungen über das Web festgestellt worden und HUBERMAN (2001, p. 25) sieht die Potenzverteilung als "… a robust empirical regularity found in all studies of the Web."
Dieser Potenzverteilung sehr ähnlich ist die Zipf-Verteilung $k \sim r^{-b}$, bei der, analog zu Zipf's Law (ADAMIC n.d.), zunächst die Knoten nach der Anzahl ihrer Kanten absteigend sortiert werden. Die Anzahl an Kanten k eines Knoten hängt nun antiproportional von seinem Rangplatz r ab, wobei b einen Wert nahe 1 annimmt. In diesem Falle ist der Vernetzungsgrad k einer Site umgekehrt proportional zu ihrem Rang.
Gemeinsam sind allen Potenzverteilungen, und damit auch den Zipf-Verteilungen, drei für die wirtschaftliche Praxis bedeutsame Eigenschaften: a) durch die extreme Schiefverteilung ist der Mittelwert bedeutungslos, b) durch die Winner-Takes-All-Verteilung ist es für jede einzelne Site sehr unwahrscheinlich, zu den Top-Sites zu gehören und c) die Verteilungen sind skalenfrei, d.h. die Verteilung weist für jede Teilmenge eine ähnliche Gestalt auf, was uns dazu ermutigte folgende explorative Studie durchzuführen.

3 Analyse der Web-Hitparade von Land24

Datengrundlage unserer Untersuchung bildeten die Ergebnisse der Land24-Hitparade (www.land24.de), an der Websites mit landwirtschaftlichen Inhalten teilnehmen und ihre Seitenaufrufe zählen lassen können. Erfaßt werden jeweils die Seitenaufrufe des aktuellen Tages, des vorigen Tages, der Vorwoche und der letzten sieben Tage. Diese Daten wurden von uns im Zeitraum vom 22.11.02 bis zum 03.02.03 an 25 Tagen erhoben und ausgewertet. Die Anzahl der Teilnehmer

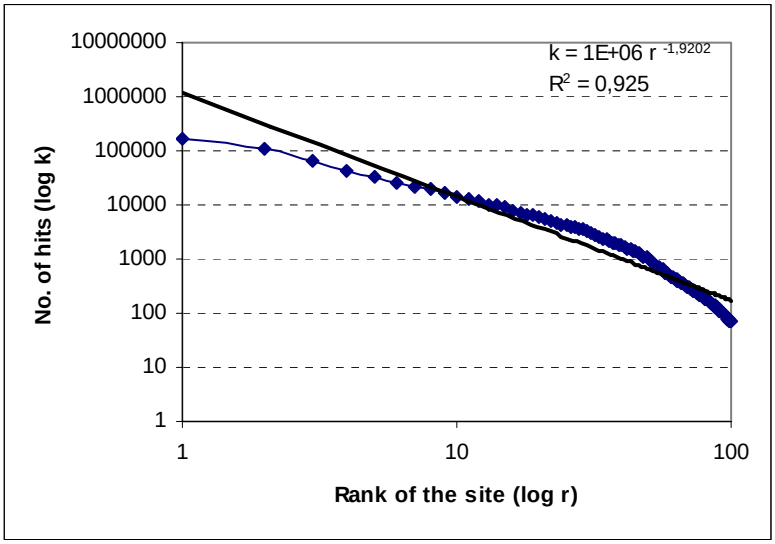

Abbildung 1: Seitenaufrufe im Verhältnis zum Rangplatz

schwankte in diesem Zeitraum zwischen 109 und 122 Websites.

Date	b	R^2	Date	b	R^2
22.11.2002	2,0063	0,9211	17.01.2003	1,9252	0,9208
27.11.2002	1,9007	0,9374	20.01.2003	1,8975	0,9285
02.12.2002	1,9569	0,9090	21.01.2003	1,8967	0,9252
03.12.2002	1,9700	0,9181	22.01.2003	1,9082	0,9268
04.12.2002	1,9695	0,9131	27.01.2003	1,9025	0,9145
05.12.2002	1,8782	0,9387	28.01.2003	1,8968	0,9280
06.12.2002	1,9877	0,9228	03.02.2003	1,9739	0,9013
17.12.2002	1,9265	0,9216	04.02.2003	1,8723	0,9260
07.01.2003	1,9387	0,9178	06.02.2003	1,8973	0,9210
09.01.2003	1,9297	0,9245	07.02.2003	1,9183	0,9174
10.01.2003	1,9528	0,9164	10.02.2003	1,8657	0,9332
15.01.2003	1,9626	0,9055	11.02.2003	1,8263	0,9340
16.01.2003	1,9463	0,9200			

Tabelle 1: Zipf-Koeffizienten b und R^2 der täglichen Daten

In der weiteren Analyse wurden die Daten nach der Anzahl der Besuche des Vortages, da diese im Gegensatz zu den Daten des aktuellen Tages immer volle 24 Stunden umfassen, sortiert und der Rangplatz der Top-100-Sites mit der Anzahl ihrer Seitenaufrufe ins Verhältnis gesetzt. Mit Hilfe von MS-Excel wurden die Parameter a und b der logarithmierten Zipf-Verteilungsfunktion $\log k_r = a - b\, r + u$ sowohl für alle 25 Einzeltage als auch für die kumulierten Daten geschätzt. Die Akkumulation der Daten erfolgte, indem alle Seitenaufrufe für jeden Platz k über alle 25 Tage aufaddiert und zur besseren Vergleichbarkeit mit den Tagesdaten durch die Anzahl der Tage dividiert wurden.

Abbildung 1 zeigt das Verhältnis zwischen der logarithmierten Anzahl an Besuchen einer Website zu ihrem logarithmierten Rangplatz für die kumulierten Daten. Der Graph bestätigt Zipf's-Law auf eindrucksvolle Weise.
Die Trendgerade ist gut an die empirischen Daten angepaßt ($R^2 = 0{,}925$) und auch der Exponent b liegt mit 1,9202 im erwarteten Bereich.
Die enorme Schiefverteilung wird am besten anhand einiger Zahlen deutlich: die Site mit den meisten Seitenaufrufen kann 24 Prozent aller Seitenaufrufe für sich verbuchen, die Top 5 kommen zusammen auf 60 Prozent und die Top 10 vereinen fast ¾ (74 Prozent) aller Besuche auf sich. Des weiteren fällt ein dreigeteilter Verlauf des Graphen auf, der im mittleren Bereich ($10 < r < 64$) oberhalb und im oberen ($r < 10$) und unteren ($r > 64$) Bereich unterhalb der Schätzgeraden verläuft; die Extrema werden also etwas überschätzt, der Mittelbereich dagegen etwas unterschätzt.
Die graphische Darstellung der 25 Tagesdaten ergibt nahezu identische Abbildungen, was aufgrund der unterstellten Eigenschaft der Skalenfreiheit auch zu erwarten war. Tabelle 1 zeigt die Bestimmtheitsmaße (R^2) und Zipf-Koeffizienten (b) für alle 25 Beobachtungstage. Bemerkenswert ist, daß die R^2 (b-Werte) nur minimal mit einer Standardabweichung von 0,0092 (0,0426) um den Mittelwert von 0,9217 (1,9243) schwanken.

4 Diskussion

Das bei Potenzverteilungen erwartete Winner-Takes-All-Phänomen wurde auch anhand dieser Daten bestätigt. Eine kleine Anzahl an Websites können eine überwiegende Mehrheit an Seitenaufrufen für sich

Share of sites [%]	Sites visited by AOL users (Hubermann 2001)			Land24 sites
	All Sites	Adult sites	edu-domain sites	
	% of visits			
0,1	32,36	1,40	2,81	-
1	55,63	15,83	23,76	24,08
5	74,81	41,75	59,50	60,21
10	82,26	59,29	74,48	74,36
50	94,92	90,76	96,88	97,84

Tabelle 2: Vergleich der Ergebnisse

verbuchen. Unklar ist allerdings, ob sich die hohe Anzahl an Besuchen auch in einer wirtschaftlichen Überlegenheit wie-derspiegelt, da der Besuch einer Seite in der Regel noch zu keinen Einnahmen des Seitenbetreibers führt.

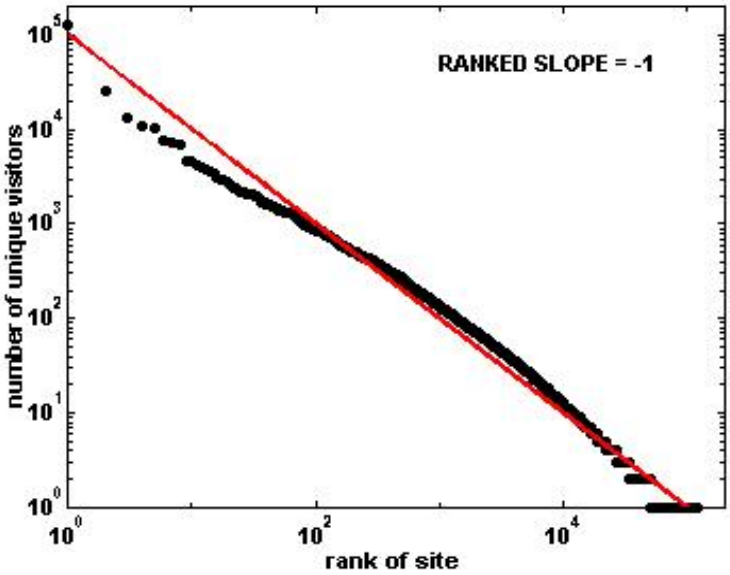

Abbildung 2: Verteilung der Seitenaufrufe von AOL-Usern (ADAMIC n.d.)

Ein Vergleich unserer Ergebnisse, die auf ca. 115 landwirtschaftlichen, deutschen Websites beruhen, mit den Resultaten von ADAMIC UND HUBERMAN (2001), die mehr als 120.000 von AOL-Benutzern aufgerufene Sites untersucht haben, zeigt zudem die Hypothese der Skalenfreiheit bestätigt. Die Daten in Tabelle 2 zeigen eine große Übereinstimmung unserer Land24-Resultate mit den Ergebnissen der „edu-domain-sites“ und auch der Graph dieser Untersuchung (Abbildung 2) weist den selben dreigeteilten Verlauf auf.
Wir hoffen anhand dieses kleinen Beispiels gezeigt zu haben, daß auch in der agrarökonomischen Forschung durch Anwendung innovativer Methoden wie der Netzwerk- oder Graphentheorie neue Erkenntnisse erzielt werden können. Gerade auf dem Forschungsgebiet der Agrarmärkte bieten sich diese Methoden an, da auch hier, wie auf allen Märkten, netzwerkartige Strukturen zu finden sind.

5 Literatur

ADAMIC, L. A. (n.d.): Zipf, power-laws, and Pareto - a ranking tutorial. XEROX Xerox Palo Alto Research Center: Palo Alto, CA. http://ginger.hpl.hp.com/shl/papers/ranking/ranking.html (21.03.2003).

ADAMIC, L.A. UND HUBERMAN, B.A. (2001): The Web's Hidden Order. XEROX PARC: Palo Alto, CA.

ALBERT, R. UND BARABÁSI, A.-L. (2001): Statistical Mechanics of Complex Networks. School of Mathematics, University of Minnesota, Minneapolis, and Department of Physics, University of Notre Dame, IN. arXiv:cond-mat/0106096v1 6Jun2001.

BARABÁSI, A.-L. (2001): The Physics of the Web. in : phyiscsweb, http://www.physicsweb.org/article/world/14/7/09 (27.01.2003).

BARABÁSI, A.-L. (2002): Linked. Cambridge, Mass.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STAT, R., TOMKINS, A. AND WIENER, J. (2000): Graph structure of the web. http://www9.org/w9cdrom/160/160.html (24.03.2003).

HUBERMAN, B.A. (2001): The Laws of the Web - Patterns in the Ecology of Information. Cambridge, Mass.

KLEINBERG, J. UND LAWRENCE (2001): The structure of the Web. Science 294 (30. Nov. 2001), S. 1849-1850.