Ein Index zur Berechnung von Prestige in Koautornetzwerken

Thomas Metz, Universität Freiburg*

5. März 2012

Zusammenfassung

Bei der Analyse von Koautornetzwerken stellt sich oft die Frage, ob sich für ein Netzwerk oder eine isolierte Komponente darin entscheiden lässt, welche Autoren besonders bedeutsam sind. Dieses Papier erläutert eine einfache Erweiterung des PageRank-Index, welche neben der Position im Netzwerk auch Kantengewichte sowie die Menge Publikationen als externe Größe nutzt. Der Index nutzt damit die in Koautornetzwerken zur Verfügung stehende Information gut aus.

Entwurf. Bitte nur nach Rücksprache zitieren.

Koautornetzwerke stellen ein wichtiges Mittel der bibliometrischen Analyse dar, sie werden aber auch als eine Form sich selbst organisierender sozialer Netzwerke angesehen [2]. Will man in einem Koautornetzwerk die Wichtigkeit eines Autors bestimmen, bieten sich mehrere Indizes an (für eine Übersicht, siehe [6, 5]). Einer davon ist der PageRank-Index, der zuerst von Brin und Page vorgestellt wurde [3]. Wir zeigen, wie sich ein an die Logik des PageRank anschließender Index aus vier Forderungen an einen „wichtigen" Forscher herleiten und anschaulich als „Wanderpokal" interpretieren lässt. Der Index nimmt die Bewertung allein aufgrund bibliometrischer Informationen vor und ist damit gut für Koautornetzwerke geeignet. Für gewöhnlich fallen bei der Erstellung eines Koautornetzwerks neben der reinen Kantenstruktur auch an, auf wie vielen gemeinsamen Publikationen eine Kante basiert sowie welche Publikationen ein Autor neben seinen Kopublikationen noch veröffentlicht hat. Ein Index der Wichtigkeit sollte diese Information möglichst vollständig nutzen.

Intuitiv lassen sich vier Kriterien fordern, die einen wichtigen Wissenschaftler auszeichnen. Ein wichtiger Wissenschaftler sollte demnach

1. gut vernetzt sein, sprich: er sollte viele Kontakte zu Kollegen haben;

*Kontakt: thomas.metz@politik.uni-freiburg.de

2. vor allem Kontakte zu wichtigen Wissenschaftlern haben;

3. vor allem Kontakte haben, die wissenschaftlich produktiv sind;

4. einem interessierten Publikum allgemein gut sichtbar sein.

Alle vier Aspekte lassen sich erfassen, wenn wir einen Index als stationäre Verteilung einer Markov-Kette konzipieren, die einen gewichteten Random Walk auf dem Koautornetzwerk mit zufälligen Sprüngen zwischen den Knoten kombiniert. Für die Herleitung folgen wir in den nächsten Abschnitten der Argumentation bei Newman [6, S. 168-178] und erweitern diese anschließend.

Forderung 1 Unsere erste Forderung – eine hohe Zahl von Kontakten – lässt sich über den Grad des Knotens erfassen. Diese Maßzahl ist als Gradzentralität bekannt [[6, S. 168]; [4, S. 337]]. Ein Forscher ist demnach umso wichtiger, je mehr Kontakte zu anderen er besitzt.

Forderung 2 Unsere zweite Forderung – Kontakte zu *wichtigen* Wissenschaftlern – kann in die Gradzentralität integriert werden, wenn man den Grad nicht als feste Eigenschaft des Knotens betrachtet sondern als Summe von abstrakten „Prestige-“ bzw. „Wichtigkeitspunkten“, die er von seinen Nachbarn erhält [6, S. 169]. Erlaubt man zusätzlich, dass Knoten die Wichtigkeit ihrer Nachbarn nicht nur sammeln sondern auch weitergeben, werden Kontakte zu wichtigen Knoten belohnt. Bezeichne $N(i)$ die Nachbarschaft und w_i die Wichtigkeit von Knoten i, lässt sich schreiben [6, S. 169]:

$$w_i = \sum_{j \in N(i)} w_j.$$

Diese Maßzahl ist bekannt als Eigenvektor-Zentralität. Ausgehend von einer zufälligen Verteilung an „Wichtigkeitspunkten“ (z.B. 1 für jeden Knoten) kann w_i iterativ geschätzt werden:

$$w' = Aw,$$

wobei der Vektor w die Wichtigkeitswerte w_i der Knoten enthält und A die Adjazentmatrix des Netzwerks bezeichnet, in der für alle Forscherpaare i, j vermerkt ist, ob eine Kante zwischen ihnen vorliegt ($a_{ij} = 1$) oder nicht ($a_{ij} = 0$; $a_{ii} = 0$). Mit zunehmender Iteration wird w proportional zum dominanten Eigenvektor von A und ein Knoten ist wichtig, wenn er viele Kanten, Kanten zu wichtigen Knoten oder beides hat [6, S. 169f.].

Gegen diese Definition lässt sich einwenden, dass Knoten mit wenigen Kanten an einen Nachbarn mehr von ihrer Wichtigkeit weitergeben sollten als Knoten mit vielen Kanten [6, S. 175]. Anders formuliert: Eine Kooperation sollte mehr Bedeutung haben, wenn sie nicht „nur eine unter vielen“ ist. Eine

entsprechende Korrektur der Eingenvektor-Zentralität nimmt der PageRank-Index vor, den Google zur Sortierung seiner Suchergebnisse verwendet [[3]; [6, S. 176]]. Bei seiner Berechnung geben Knoten einen Anteil ihrer Wichtigkeit proportional zur Anzahl ihrer Kanten weiter. Formal:

$$w_i = \sum_{j \in N(i)} \frac{1}{|N(j)|} w_j.$$

Der PageRank-Index kann auch für große Netzwerke einfach berechnet werden. Hierzu normiert man die Einträge in A auf die Summe über alle Einträge im entsprechenden Spaltenvektor. Formal:

$$h_{ij} = \frac{a_{ij}}{\sum_i a_{ij}}.$$

Die Einträge der so entstehenden stochastischen Matrix H sind nichtnegativ und summieren spaltenweise zu 1. Gegeben einen Vektor w der Wichtigkeit der einzelnen Konten (ebefalls normiert auf die Summe 1) lässt sich die Berechnung des PageRank-Index formulieren als [1]: $w = Hw$. Mit anderen Worten: Der Vektor an Wichtigkeitswerten ist ein Eigenvektor von H zum Eigenwert 1 oder genauer: der stationäre Eigenvektor von H.

Die Normierung zu H eröffnet eine Interpretation der Matrix als stochastischen Kern einer Markov-Kette, die einen Random Walk auf dem Graphen beschreibt [1]. Damit wird auch der PageRank-Wert eines Knotens als stationäre Verteilung von H anschaulich lesbar: Folgt man mit dem Finger den Kanten des Netzwerks und entscheidet an jedem Knoten zufällig aufs Neue, welcher Kante man folgt, gibt der Eintrag h_{ij} wider, mit welcher Wahrscheinlichkeit man von j nach i wandert. In der stationären Verteilung ist der Wert w_i dann der Anteil der Zeit, die man auf Knoten i verbringen würde, wenn man lange genug über das Netzwerk läuft. Als einfache Methode der Berechnung bietet sich die Potenzmethode an.[1]

Analog lässt sich Wichtigkeit somit als eine Art „Wanderpokal“ vorstellen, der jede Runde von Forscher zu Forscher weitergereicht wird. Über einen langen Zeitraum hinweg beobachtet würde sich der Pokal vor allem bei Forschern aufhalten, die entweder selbst viele Kontakte haben oder die Kontakte zu wichtigen Forschern haben – letztere würden selbst den Pokal oft erhalten und ihn deshalb auch oft weiterreichen können.[2]

[1] Die Potenzmethode besteht darin, den Vektor x, dessen Elemente zu 1 summieren, iterativ mit dem Markov-Kern P zu multiplizieren. Für diese Methode kann man zeigen, dass der Vektor – oft nach nur wenigen Iterationen – zum Eigenvektor des größten Eigenwerts und damit zum von uns gesuchten Vektor der Wichtigkeitswerte) konvergiert [1]. Formal: $x^{t+1} = Px^t$. Versteht man die ursprünglichen Elemente des Vektors als Wahrscheinlichkeit, mit dem Finger auf einen bestimmten Knoten zu zeigen und setzt alle Elemente des Vektors auf 0 bis auf eines, das den Wert 1 erhält (auf diesen Knoten zeigt man zu Beginn), erschließt sich in den ersten Multiplikationen das Verfahren sehr intuitiv.

[2] Im ursprünglichen Anwendungskontext des PageRank-Index ist die äquivalente Interpretation die eines Surfers, der sich zufällig durchs Internet bewegt, indem er wahllos auf

Forderung 3 Die Interpretation des PageRank-Index ist bereits sehr anschaulich, allerdings fehlt noch unsere dritte Forderung – nämlich, dass sich Wichtigkeit vor allem aus produktiven Kontakten herleitet. Dies kann in den PageRank-Index aufgenommen werden, wenn man bevorzugt „dicken" Kanten folgt, i.e. Kanten, die ein höheres Gewicht haben. Sei L eine Matrix, deren Einträge l_{ij} die Kantengewichte (in unserem Fall: die Zahl der Kopublikationen von i und j) enthalten, kann man per Normierung analog zu H eine stochastische Matrix M berechnen, deren Übergangswahrscheinlichkeiten zusätzlich berücksichtigen, welches relative Gewicht die Kante an allen ausgehenden Kanten eines Knoten besitzt:

$$m_{ij} = \frac{l_{ij}}{\sum_i l_{ij}}$$

Der Vektor der Wichtigkeitswerte w berechnet sich weiterhin analog als $w = Mw$ und lässt sich ebenfalls als „Wanderpokal" interpretieren, der nun aber bevorzugt an Kollegen weitergereicht wird, mit denen man viel zusammenarbeitet.

Unser bisheriger Index ist für ein Netzwerk bzw. dessen Komponenten problemlos per Potenzmethode zu berechnen, da der Graph ungerichtet und damit die unterliegenden Matrizen sowohl irreduzibel (wegen der ungerichteten Kanten kann der „Wanderpokal" nicht in einen Teil des Netzes geraten, aus dem er nicht mehr heraus kommt) als auch primitiv (da keine gerichteten Kanten vorliegen, ist es ist möglich, einen Knoten in der Periode 1 zu erreichen) sind. Für ungerichtete Graphen stellt sich jedoch das Problem, dass die resultierenden Wichtigkeitswerte für den PageRank-Index proportional zum Grad des Knoten [[5, S. 43]; [6, S. 177]] bzw. – für unsere Erweiterung – proportional zur Summe der Kantengewichte eines Knoten sind. Dieses Problem lässt sich relativieren, wenn man unsere vierte Forderung berücksichtigt.

Forderung 4 Unsere vierte Forderung besagt, dass ein wichtiger Forscher auch von außen sichtbar sein sollte. Dies lässt sich über die Gesamtzahl der Publikationen erfassen – je mehr es davon gibt, desto sichtbarer ist unser Forscher. Zugleich besagt diese Forderung, dass Wichtigkeit nicht allein aus den Verbindungen zu anderen Forschern entsteht sondern auch an einer externen Quelle stammen kann, nämlich der Länge der Publikationsliste.

Eine einfache Art, eine externe Quelle von Wichtigkeit zu modellieren, ist die Einführung eines Terms, der Knoten eine Bedeutung „qua Existenz" einräumt – sei es als konstanter Beitrag oder spezifisch für jeden Knoten getrennt [6, S. 177]. Dieser Term wird meist verwendet, um zu verhindern, dass in einem gerichteten Netzwerk ein Bereich entsteht, der unseren „Wanderpokal" zwar aufnimmt, aus dem heraus aber keine Verbindungen mehr

Links auf Webseiten klickt. Seiten, auf denen er in seiner Reise am meisten Zeit verbringt, sind potenziell wichtig und erhalten einen hohen Indexwert [3, 1].

nach außen weisen, wodurch die Knoten in dem entsprechenden Bereich alle Wichtigkeit aufnehmen würden [1] bzw. dass Knoten ohne eine eingehende Verbindung eine Wichtigkeit von 0 erhalten, die sie dann an andere weiterreichen [6, S. 171f.]. Aus unserer Warte stellt die Aufnahme einer externen Quelle unter anderem sicher, dass die vorhandene bibliographische Information über einen Autor optimal genutzt wird.

Da wir von einer stochastischen Matrix M ausgehen, welche die netzwerkinterne Wahrscheinlichkeit umfasst, von einem Autor zum nächsten zu wechseln, liegt es nahe, die externe Information in analoger Weise einzubeziehen. Wir erreichen dies, indem wir eine Matrix S definieren, deren Elemente $s_{i\cdot}$ verzeichnen, wie groß der Anteil der Publikationen von Autor i an allen Publikationen im Netzwerk ist. Mit anderen Worten: S besteht aus Spaltenvektoren, welche die Wahrscheinlichkeit enthalten, bei Auswahl einer zufälligen Publikation des einem Kooperationsnetzwerk unterliegenden Datensatzes Autor i als Alleinautor oder als Koautor vorzufinden. In Analogie zur Interpretation eines Random Walk würde S ein zufälliges Springen zwischen den Autoren erfassen, das proportional ist zu ihrem Output an Arbeiten – gleich einem Leser, der in der Bibliothek die von uns erfassten Zeitschriften durchblättert, um zufällig einen interessanten Artikel auszuwählen. Darauf aufbauend können wir eine Matrix P als gewichtetes Mittel berechnen, die beide Informationen mischt:

$$P = \alpha M + (1 - \alpha)S.$$

Interpretativ bedeutet sie, dass der „Wanderpokal“ nicht mehr nur alleine den Kanten des Netzwerks folgt sondern mit einer Wahrscheinlichkeit von α zufällig zu einem Knoten springt – wobei er jene Autoren bevorzugt, die viele Publikationen haben (siehe auch [1]). Die Erüllung von Forderung 4 relativiert das oben diskutierte Problem insofern, als dass eine unabhängige Informationsquelle hinzugezogen wird. Solange nicht alle Autoren die gleiche Zahl an Publikationen besitzen springt der „Wanderpokal“ einzelne Knoten öfter an, die dadurch mehr Prestige weiterreichen können als ihnen durch ihre bloße Position im Netzwerk möglich wäre.

Prinzipiell ist die Wahl von α bis auf die Extreme von 0 (Ignorieren des Netzwerks) und 1 (Trivialisierung des Index) freigestellt, in Anbetracht der Tatsache, dass eine externe Quelle von Wichtigkeit jedoch eine von insgesamt vier Forderungen an unseren Index ist, erscheint es sinnvoll, durch einen Wert von $\alpha = 0,75$ sicherzustellen, dass die „Sprünge“ unseres „Wanderpokals“ einen Anteil von 25 Prozent an P haben. Da Knoten ohne Kanten in M einen Spaltenvektor der Summe 0 produzieren würden, müssen wir Autoren ohne Kopublikationen von der Berechnung ausschließen.

Literatur

[1] David Austin. How google finds your needle in the web's haystack. *American Mathematical Society Feature Column*, 10(12), December 2006. online: http://www.ams.org/samplings/feature-column/fcarc-pagerank.

[2] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vícsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590–614, 2002.

[3] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107 – 117, 1998. Proceedings of the Seventh International World Wide Web Conference.

[4] Alessio Cardillo, Salvatore Scellato, and Vito Latora. A topological analysis of scientific coauthorship networks. *Physica A. Statistical Mechanics and its Applications*, 372(2):333–339, 2006.

[5] Dirk Koschützki, Katharina Lehmann, Leon Peeters, Stefan Richter, Dagmar Tenfelde-Podehl, and Oliver Zlotowski. Centrality indices. In Ulrik Brandes and Thomas Erlebach, editors, *Network Analysis*, volume 3418 of *Lecture Notes in Computer Science*, pages 16–61. Springer, Berlin, 2005.

[6] M.E.J. Newman. *Networks. An Introduction.* Oxford University Press, Oxford, 2010.