Mathematik hinter Google

Zulassungsarbeit

Eberhard-Karls-Universität Tübingen
Mathematisches Institut

Arbeitsbereich Funktionalanalyis

Prof. Dr. Rainer Nagel
Prof. Dr. Ulf Schlotterbeck

Eingereicht von

Matthias Frick

Tübingen, im Juni 2007

Erklärung

Ich erkläre, dass ich die Arbeit selbständig angefertigt und nur die angegebenen Hilfsmittel benutzt habe. Alle Stellen, die dem Wortlaut oder dem Sinn nach anderen Werken, gegebenenfalls auch elektronischen Medien, entnommen sind, sind von mir durch Angabe der Quelle als Entlehnung kenntlich gemacht. Entlehnungen aus dem Internet sind durch Ausdruck belegt.

Tübingen, den 27. Juni 2007

Danksagung

Mein herzlicher Dank gilt in erster Linie Prof. Dr. Rainer Nagel und Prof. Dr. Ulf Schlotterbeck, die mich in die Arbeitsgruppe Funktionalanalysis aufnahmen und mich dazu ermutigten, meine Zulassungsarbeit im Fach Mathematik anzufertigen. Ich danke ihnen für die sehr engagierte Betreuung, die vielen Anregungen beim Entwickeln der Gedanken und nicht zuletzt für die tolle Atmosphäre, die sie durch ihre verständnisvolle und freundliche Art in der Arbeitsgruppe stifteten.

Darüber hinaus gilt mein Dank auch allen anderen Mitgliedern der Arbeitsgruppe, die nicht weniger für das angenehme Arbeitsklima verantwortlich sind und mir bei allen mathematischen Fragen hilfsbereit und mit großer Geduld zur Seite standen.

Einen ganz besonderen Dank aussprechen möchte ich auch dem ehemaligen Führungsduo der Fußball-Hobbymannschaft Real Analysis, Thomas Stumpp und Tobias Jahnke, die mir ein glühendes Beispiel dafür waren, dass man auch als begeisterter Fußballer ein erfolgreiches Mathestudium absolvieren kann, und mich in Zeiten des Zweifelns und des Frustes während des Grundstudiums immer wieder neu motivierten.

Last but not least danke ich meinen Freunden und allen voran meinen Eltern, die mich während meines gesamten Studiums und beim Erstellen dieser Arbeit tatkräftig unterstützt haben.

Inhaltsverzeichnis

1 Einleitung

1.1 Generation Google

Das Informationsbeschaffungs-Verhalten in unserer Gesellschaft hat sich in den letzten Jahren dramatisch verändert. Informationen, die man früher über Bücher, Zeitschriften, Fernsehen, Radio oder gar persönliche Gespräche einholte, sind heute – dank Internetsuchmaschinen – nur einen „Mausklick“ entfernt. Beeindruckend ist nicht nur der so einfache Zugang zu beliebigen Informationen, sondern auch die rasante Geschwindigkeit, mit der sich diese mediale Veränderung vollzogen hat. Vergleicht man den Zeitraum, in dem sich Internet und Online-Suche seit ihren Geburtsstunden ausgebreitet haben, mit der Zeit, die es einst dauerte, bis die ersten gedruckten Bücher für die breite Masse erhältlich waren, so wirkt die Nutzung von Suchmaschinen fast wie ein größerer Durchbruch als Gutenbergs Buchdruck.

Heute, nur knapp 9 Jahre nach Googles Online-Gang im September 1998, scheint es für uns selbstverständlich, dass die Unmengen an Informationen, die das World Wide Web bietet, fast uneingeschränkt zugänglich sind. Doch dies ist keineswegs selbstverständlich und ausschließlich den Suchmaschinen-Entwicklern zu verdanken. Das Internet allein ist eigentlich nichts weiter als eine ungeordnete Ansammlung an unzähligen Daten, die einem ständigen, dynamischen Wachstum und Wandel unterliegt. Eine gezielte Suche nach Informationen zu beliebigen Stichwörtern wäre ohne Suchmaschine nicht möglich. Suchmaschinen bringen Ordnung ins Chaos des World Wide Web und machen das Internet zu dem, was es für den Nutzer heutzutage ist: Eine schier unerschöpfliche Quelle an Informationen.

Die erfolgreichste aller Suchmaschinen ist zweifellos Google: Knapp 40 Prozent Marktanteil bei Suchmaschinenanfragen weltweit ([LaMey], Seite 4) und sogar über 90 Prozent bei deutschen Internet-Suchen [FN07] sprechen eine deutliche Sprache: Wer im Internet surft, kommt an Google kaum noch vorbei. Diese Einsicht hatte wohl auch die Duden-Redaktion, als sie 2005 das populäre Verb *googeln* als Synonym für das allgemeine Suchen im Internet (also auch mit anderen Suchmaschinen) aufnahm[1]. Kein Wunder also, dass oft von der *Generation Google* die Rede ist, wenn man die heutige Gesellschaft des Internet-Zeitalters beschreibt.

Schon heute ist Google mit einem Börsenwert von über 120 Milliarden Euro unter den drei erfolgreichsten IT-Unternehmen der Welt. Das Handelsblatt bezeichnete Google-Chef Eric Schmidt 2006 als Mann des Jahres [HB06], die Google-Gründer Larry Page und Sergey Brin sind längst Multimillionäre und dennoch ist noch kein Ende der Erfolgsstory in Sicht.

[1] Mittlerweile wurde der Eintrag – interessanterweise aus Markenschutzgründen auf Veranlassung der Google-Betreiber selbst – in „mit Google im Internet suchen“ geändert.

Google wächst und expandiert im Eiltempo. Immer mehr Dienstleistungen wie *Google Maps*, *Google Mail* oder *Google Video* werden im World Wide Web angeboten (siehe Abb. 1.1). Erst im Oktober 2006 wurde das Internet-Videoportal YouTube kurzerhand für 1,65 Milliarden Dollar übernommen [Goog2] und im April 2007 machte Google gar für stolze 3,1 Milliarden das Rennen um die Übernahme der umworbenen Online-Werbefirma Doubleclick und stach dabei den Konkurrenten Microsoft aus.

- Suchmaschine *http://www.google.de*
- Google Mail *http://mail.google.com/*
- Google Groups *http://groups.google.de/*
- Google News *http://news.google.de/*
- Google Maps *http://maps.google.de/*
- Google Earth *http://earth.google.de/*
- Google Video *http://video.google.de/*
- Google Translate *http://www.google.de/language_tools*
- Google Scholar *http://scholar.google.de/*
- Google Books *http://www.google.de/books*
- Froogle *http://www.google.de/froogle*
- ...

Abbildung 1.1: Internet-Dienstleistungen von Google

Trotz dieses mittlerweile breitgefächerten Dienstleistungsangebots der Marke *Google* (siehe auch [Goog1]) basiert ihr heute fast unglaublicher Börsenwert auf dem ursprünglichen Erfolg und der nach wie vor marktführenden Position der Suchmaschine.

„Google zeigt mich, also bin ich“ [Kaim]. So lautet nicht nur der Titel eines 45minütigen französischen Kurzfilms über die Abhängigkeit Jugendlicher vom Internet und deren Lebensgefühl, sondern auch die Einsicht, der sich weltweit immer mehr Unternehmen stellen. Jeder buhlt um die vordersten Plätze bei den Google-Suchergebnissen. In einer Welt, in der sich auch das Geschäftsleben mehr und mehr online vollzieht, ist es für Firmen und Dienstleister immer wichtiger, von potentiellen Kunden im Internet gefunden zu werden. Angesichts der Monopolstellung Googles und des Wettkampfs um die vorderen Plätze der Suchergebnis-Listen scheinen Aussagen wie „eine Firma, die man bei Google nicht findet, existiert nicht“ [FN07] nicht vermessen.

Google ist also in aller Munde, und – noch viel wichtiger – immer häufiger in fast allen unseren Browserfenstern. Es ist also von aktueller Relevanz, sich zu fragen, worauf dieser unglaubliche Erfolg beruht. Welches Erfolgsgeheimnis steckt hinter Googles Monopolstellung? Eine mögliche Antwort lautet: Jede Menge Mathematik!
Im Hauptteil dieser Arbeit soll diese *Mathematik hinter Google* beleuchtet werden. Zuvor jedoch werden einige grundsätzliche Dinge über die Funktionsweise von Suchmaschinen erläutert, um verständlich zu machen, wo die Mathematik ansetzt und warum sie hier so wichtig ist.

1.2 Wie funktioniert eine Suchmaschine?

Suchmaschinen machen uns die unfassbaren Mengen an Internet-Informationen erst zugänglich. Suchmaschinen öffnen uns – bildlich gesprochen – die Türen zur Internet-Bibliothek und den Informations-Lagern zu bestimmten Suchbegriffen, sie sortieren die Informationen zusätzlich nach Relevanz und lassen uns so ohne Zeitverlust gleich beim Eintritt auf die (angeblich) relevantesten Informationen stoßen. Das ist eine beachtliche Leistung, denn es gibt mehrere Milliarden Webseiten im World Wide Web und dazu fast ebenso viele Suchanfragen pro Tag. Wie können sich die Suchmaschinen in diesem inhomogenen und immer noch rasant wachsenden und sich ständig verändernden Netz auskennen und der unglaublichen Nachfrage gerecht werden?

Alle Suchmaschinenbetreiber benutzen hierzu zunächst eine automatische Surfsoftware (Spider, Crawler, Webrobots oder kurz Bots), die permanent und immer wieder aufs Neue das sich ständig verändernde und wachsende Internet durchforstet. Das Surf-Programm nutzt hierbei die Hyperlinks zwischen den Webseiten und hangelt sich von Seite zu Seite und von Hyperlink zu Hyperlink, um so viele Seiten mit öffentlichem Zugang wie möglich zu orten. Mit den angesteuerten Seiten geschieht zunächst einmal Dreierlei:

1. Jede Seite bekommt eine Nummer.
2. Eine Kopie der Seite wird im firmeneigenen Rechenzentrum gespeichert.
3. Der Inhalt des gesamten Quelltextes der Seite (d.h. sowohl der Fließtext als auch Titel, Meta-Tags, Keywords, Anchortext...) wird analysiert und in ein riesiges Schlagwort-Verzeichnis – den sogenannten Index – abgelegt und sortiert.

Der Index ist vergleichbar mit einem Stichwortverzeichnis am Ende eines Buches. Neben jedem Begriff sind die Nummern all derjenigen Seiten aufgelistet, auf denen der Begriff zu finden ist.

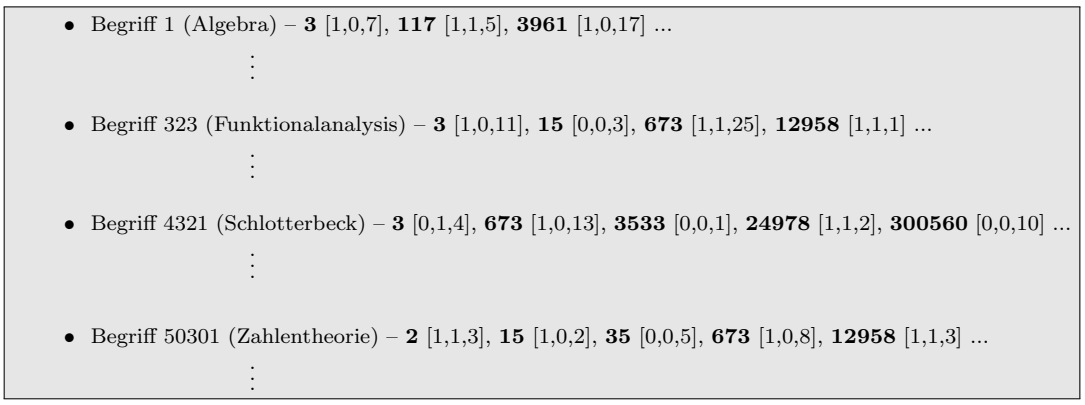


Abbildung 1.2: Vereinfachtes Modell-Beispiel für einen Suchmaschinen-Index (vgl. [LaMey], Kapitel 2.2)

In Abbildung 1.2 stehen die fettgedruckten Zahlen für die Seiten, auf denen der jeweilige Begriff zu finden ist. Die Vektoren in eckigen Klammern dahinter geben quantitative

und qualitative Auskünfte über die Art und Weise der Verwendung des Begriffs auf der Seite, also z.B. den genauen Ort (Titelzeile, Keywords, Fließtext...) oder die Häufigkeit des Vorkommens. Ferner kann hier beispielsweise vermerkt werden, ob der Begriff durch besonderes Layout hervorgehoben wird (unterstrichen, fettgedruckt, größere Schriftart...). In diesem Beispiel steht die erste Vektorkomponente für ein etwaiges Vorkommen des Begriffs im Titel der Webseite. Hier ist eine 1 eingetragen, falls die Titelzeile den Begriff enthält, andernfalls eine 0. Die zweite Stelle gibt an, ob der Begriff in den vom Webautor angegebenen Keywords der Seite erwähnt ist (falls ja: 1, sonst: 0). Die letzte Vektorstelle steht für die Anzahl der Verwendungen des Begriffs auf der gesamten Seite.

Bei Suchanfragen wird nun geprüft, ob die eingegebenen Suchbegriffe im Index vorhanden sind. In den meisten Fällen findet die Suchmaschine hierbei mehrere tausend Treffer, weshalb eine Sortierung der Trefferliste nötig ist, um den Nutzern auf der Suche nach tatsächlich relevanten Treffern ein allzu langes Durchforsten sogenannten Info-Mülls zu ersparen. Die hierfür zuständigen Ranking-Algorithmen sind also entscheidend für den tatsächlichen Nutzen und damit den Erfolg einer Suchmaschine. Eine Möglichkeit für die Sortierung der Treffer bietet die Berücksichtigung zusätzlicher Informationen über die unterschiedliche Verwendung der Suchbegriffe auf den verschiedenen Seiten der Trefferliste. Hierbei wird versucht, die Relevanz der Seiten zum jeweiligen Suchbegriff zu beurteilen.

Im gewählten Beispiel kann also der dreistellige Vektor herangezogen werden, um jeder Seite abhängig von ihrem Inhalt bzw. dem jeweiligen Suchbegriff einen sogenannten „Content-Score“[LaMey] zuzuordnen, z.B. durch einfache Addition der Vektorstellen.

Eine Suche nach dem Begriff „Funktionalanalysis“würde hier also zunächst die Seiten 3, 15, 673 und 12958 als Treffer liefern, wobei 673 aufgrund des höchsten Content-Scores (1+1+25=27) an erster Stelle stehen würde. Oft beinhaltet eine Suchanfrage zwei oder mehrere Begriffe, z.B. „Funktionalanalysis Schlotterbeck“. Eine solche Anfrage würde in diesem Fall die Seiten 3 und 673 als relevante Treffer liefern, da auf beiden Seiten beide Begriffe vorkommen. Die von beiden Suchbegriffen abhängigen Content-Scores der beiden Seiten ergeben sich dann beispielsweise durch Multiplikation der entsprechenden Vektoren:

$$\text{Content-Score Seite 3} = \underbrace{(1+0+11)}_{Funktionalanalysis} \times \underbrace{(0+1+4)}_{Schlotterbeck} = 60$$

$$\text{Content-Score Seite 673} = \underbrace{(1+1+25)}_{Funktionalanalysis} \times \underbrace{(1+0+13)}_{Schlotterbeck} = 378$$

Natürlich werden in der Realität noch weit mehr Informationen als in unserem dreistelligen Vektor berücksichtigt und auch die Gewichtung der einzelnen Komponenten und die numerische Berechnung sind viel komplizierter. So müssen z.B. auch Faktoren wie die räumliche Nähe zweier oder gegebenenfalls mehrerer Suchterme zueinander und deren Eingabereihenfolge auf den jeweiligen Seiten berücksichtigt und mit in die Berechnung einbezogen werden. Das gewählte Modell-Beispiel soll also lediglich zur Veranschaulichung des Prinzips der Inhaltsindizierung von Webseiten dienen.

Bis zum Durchbruch von Google nutzten Suchmaschinen zur Erstellung eines Suchergebnis-Rankings jedoch fast ausschließlich derartige Suchanfragen-abhängige Content-Scores. Dies bot eine breite Angriffsfläche für einfache Manipulationen sogenannter Spammer (zum Beispiel durch Mehrfachnennung eines populären Suchbegriffs in weißer Schriftfarbe auf weißem Hintergrund) und resultierte in oft hohen Rankingwerten für qualitativ schlechte Webseiten.

1.3 Warum Google?

Google ist derzeit mit weit über 250 Millionen Suchanfragen pro Tag die mit Abstand erfolgreichste Internet-Suchmaschine. Grund genug, sich zu fragen, worin sich Google von anderen Suchmaschinen unterscheidet und was Googles Vorzüge sind. Einige mögliche Gründe für Googles Erfolg sind beispielsweise:

- *Übersichtliches Layout und einfache Bedienung*
  Im Vergleich zu konkurrierenden Suchmaschinen hob sich Google von Anfang an durch seine einfache und übersichtliche Benutzeroberfläche mit dem intuitiv bedienbaren Sucheingabe-Feld im Zentrum ab. Dies ermöglicht einerseits Such-Anfängern eine erfolgreiche Suche, bietet andererseits aber auch Könnern über zusätzliche Eingabe-Codes die Möglichkeit einer detaillierten Fortgeschrittenen-Suche.
  Auch die übersichtliche Präsentation der Suchergebnisse, bei der unter der Angabe des Titels als kleine Seitenvorschau das direkte Umfeld des Suchbegriffs auf der Seite angezeigt wird, ist sicherlich zu den Vorzügen Googles zu zählen (vgl. [Schö]).

- *Schnelle Beantwortung der Suchanfrage*
  Googles Richtwert für die Antwortzeit ist eine halbe Sekunde.

- *Größter Index*
  Seit September 2005 wird die Größe des Index nicht mehr auf der Hauptseite angezeigt, bis dahin waren schon über 8 Milliarden Internetseiten indiziert.
  Google bietet Webautoren, die eine neue Seite ins Netz stellen und die automatische Index-Erfassung durch den Google-Bot nicht abwarten wollen, die Möglichkeit, unter http://www.google.com/addurl.html ihre Seite zur Google-Liste der „to-be-crawled URLs“hinzuzufügen, um so die Erfassung der Seite zu beschleunigen.

- *Plausibles Ranking der Suchergebnisse*
  Das Google-Ranking ist das eigentliche Erfolgsgeheimnis des Unternehmens. Denn schon in den Anfangszeiten des aufblühenden Suchmaschinenmarktes war Googles entscheidender Vorteil gegenüber der Konkurrenz die offensichtlich gute Informationsfilterung, also die plausible Relevanz-Bewertung der Webseiten zur Suchanfrage: „Google always seemed to deliver the good stuff upfront“[BrLei].

Wie schon erwähnt sind letztendlich die Ranking-Algorithmen entscheidend für den Erfolg einer Suchmaschine. Dementsprechend musste der Google-Algorithmus von Beginn an besonders innovativ und gut sein.

Eine von vielen Innovationen der Google-Technologie war die Miteinbeziehung der Anchor Text Indizierung in die Berechnung des Content-Scores. D.h. Google indiziert auch die Anchor-Texte der Inlinks (Hyperlink-Descriptions) einer Seite – also eine externe Seitenbeschreibung durch andere Webautoren – um den Einfluss eigenen Inhalts-Spammings (wie oben beschrieben) zu minimieren. Dies führte zwischenzeitlich wiederum zum Problem sogenannter *Google-Bomben*, als beispielsweise zahlreiche Gegner von George W. Bush einen Link mit der Beschreibung *Miserable Failure* auf die offizielle Homepage des Weißen Hauses setzten und somit die Bush Biographie auf Rang 1 der Google-Ergebnisliste zur Suche nach *Miserable Failure* katapultierten.

Die entscheidende Neuerung, die Googles Algorithmus 1998 von der Konkurrenz unterschied, war allerdings, dass zur Relevanzbewertung einer Seite für eine bestimmte Suchanfrage nicht nur ein Suchanfragen-abhängiger Content-Score (wie in 1.2 beschrieben), sondern auch ein Suchanfragen-unabhängiger Wichtigkeitswert einer Seite („Popularity Score“[LaMey] oder „Importance Score“[BrLei]) nach einem von den Google-Gründern Larry Page und Sergey Brin entwickelten Verfahren (das sogenannte PageRank-Verfahren) berechnet wurde und mit ersterem zu einem Gesamt-Rankingwert („Overall Score“[LaMey]) kombiniert wurde.

Man könnte vermuten, dass dieser Mehraufwand zu einer deutlich langsameren Beantwortung der Suchanfrage führt. Doch dies ist nicht der Fall. Da der PageRank-Wert in regelmäßgien Abständen offline und damit insbesondere unabhängig von der Suchanfrage berechnet wird, muss er im Moment der Suchanfrage lediglich abgerufen werden. In den Bruchteilen einer Sekunde von der Suchanfrage bis zur Präsentation der Suchergebnisse prüft die Suchmaschine im Index, welche Seiten die Suchbegriffe enthalten, berechnet über die jeweiligen Begriffs-Vektoren den Content-Score dieser Seiten, kombiniert diesen mit ihrem PageRank-Wert und ordnet die Seiten nach der Reihenfolge ihrer Gesamt-Rankingwerte.

Die genaue Gewichtung der einzelnen Faktoren, also der komplette Google-Algorithmus mit seinen weit über 100 Parametern, ist natürlich das bestgehütete Firmengeheimnis Googles. Auf dem Weg zur „perfekten Suchmaschine“[Goog3] mit möglichst „optimalen Parametern“[Schö] nutzen die „dauerhaft um Innovation bemühten“[Goog3] Google-Entwickler allerdings auch heute noch gezieltes User-Feedback, um durch weitere Veränderungen der Parameter das Verhalten des Ranking-Algorithmus' immer mehr an die menschlichen Einschätzungen anzunähern.

Trotz aller Entwicklung und Geheimniskrämerei gibt Google auf seiner Homepage ganz offen preis, dass die Berechnung des Suchanfragen-unabhängigen Wichtigkeitswerts durch das revolutionäre PageRank-Verfahren nach wie vor den wichtigsten Baustein des Google-Algorithmus darstellt: „Das Herz unserer Software ist PageRank“[Goog4]. Obwohl auch hinter den anderen beteiligten Verfahren des Gesamtalgorithmus jede Menge interessante Informatik und numerische Mathematik steckt, wird sich diese Arbeit daher im Folgenden auf die Berechnung des PageRank-Wertes konzentrieren. Der Titel dieser Arbeit – *Mathematik hinter Google* – meint also genauer *Mathematik hinter PageRank*.

1.4 Die Entwicklung der PageRank-Idee

Die zündende Idee hinter PageRank war, das Internet selbst bzw. seine Linkstruktur über die Bewertung seiner Webseiten entscheiden zu lassen. Hierbei macht man sich die demokratische Natur des World Wide Web zu Nutze und interpretiert die Links, die ein Webautor – also ein Wähler in einer demokratischen Wahl – auf andere Seiten setzt, als Stimmabgabe für die verlinkten Seiten. Denn indem ein Autor auf andere Seiten verweist, drückt er gewissermaßen seine subjektive hohe Wertschätzung dieser Seiten aus. Die Gesamtheit dieser subjektiven Stimmabgaben – also die gesamte Linkstruktur – kann man als demokratische Wahl und damit als Mittel zur Bestimmung eines „globalen Bedeutsamkeitswertes“[Schö] auffassen.

Um diese Idee in Mathematik zu übersetzen und damit arbeiten zu können, sind zunächst einige Notationen und Definitionen nötig.

Notationen und Definitionen

- Sei $n \in \mathbb{N}$ die Anzahl aller Webseiten,
  $W = \{W_k \mid 1 \leq k \leq n, \quad k \in \mathbb{N}\}$ die Menge aller Webseiten.
- $I_k := \{i \mid \exists$ ein Link von der Webseite W_i zur Webseite $W_k\}$
  $=$ Menge der auf W_k verweisenden Links (Inlinks).
- $O_k := \{j \mid \exists$ ein Link von der Webseite W_k zur Webseite $W_j\}$
  $=$ Menge der von W_k ausgehenden Links (Outlinks).
- Im Folgenden bezeichne $0 \leq x_k$ den Wichtigkeitswert der Seite W_k, so dass
  $x_i > x_j \Leftrightarrow W_i$ ist wichtiger als W_j.

Erste Überlegungen

Mit dem Ansatz, die Links als Stimmabgaben zu sehen, ist klar, dass eine Webseite mit jedem Inlink – also jeder Stimme – an Wichtigkeit gewinnt. Ohne weitere Vorüberlegungen macht die Idee, die Wichtigkeit einer Seite einfach mit der Anzahl seiner Inlinks gleichzusetzen, zunächst also Sinn. Mit der eingeführten Notation hieße das also

$$x_k = |I_k|.$$

Hierbei würde jedoch die Herkunft der Links ignoriert werden: Ein Inlink von einer unbedeutenden, privaten Homepage könnte die eigene Seite in gleichem Maße aufwerten wie ein Inlink z.B. von www.yahoo.com.

Nicht nur die Anzahl, sondern auch die Qualität der Inlinks – impliziert durch die Wichtigkeit des Link-Setzers – muss also über die Wichtigkeit einer Seite entscheiden. So könnte man etwa die Wichtigkeit einer Seite W_k rekursiv durch die Summe der Wichtigkeitswerte derjenigen Seiten bestimmen, die mit einem Link auf W_k verweisen:

$$x_k = \sum_{i \in I_k} x_i$$

Dieser Ansatz stellt zwar gegenüber ersterem rein logisch eine Verbesserung dar, weil nun tatsächlich die Herkunft und damit die vermeintliche Qualität der Inlinks in die Berechnung des Wichtigkeitswertes einginge, bedarf jedoch immer noch weiterer Überlegungen und einer kleinen Modifikation.

Zunächst beißt sich hier sozusagen die Katze selbst in den Schwanz, denn der Wichtigkeitswert von W_k soll über die ihrerseits unbestimmten Wichtigkeitswerte der auf W_k verweisenden W_i bestimmt werden. Wie sollen die x_k also konkret berechnet werden? Dieses Berechnungsproblem soll jedoch erst einmal hintangestellt werden.
Viel wichtiger ist die Tatsache, dass die bisherige Formel noch nicht – wie gewünscht – demokratischen Grundsätzen entspricht: Durch Setzung eines Links könnte ein Webautor seine komplette Wichtigkeit auf den Wichtigkeitswert der verlinkten Seite übertragen. Da jeder Webautor beliebig viele Links im World Wide Web setzen kann, könnten besonders fleißige Autoren also unbegrenzten Einfluss in unserer Web-Popularitätswahl ausüben. Anders ausgedrückt: Die Wähler hätten beliebig viele Stimmen in diesem Wahlverfahren! Eine demokratische Formel müsste diesen unbeschränkten Einfluss also irgendwie limitieren. Lawrence Page und Sergey Brin setzten daher die folgende Interpretation in ihrer Formel um.

Der PageRank-Ansatz

Die PageRank-Formel gewährleistet, dass die gesamte Linksetzung eines Webautors als nur eine Stimme interpretiert werden kann, die durch die verschiedenen Links zu gleichen Teilen aufgesplittet wird.

$$x_k = \sum_{i \in I_k} \frac{x_i}{|O_i|} \tag{1.1}$$

Jeder Summand x_i (Wichtigkeitswert von W_i) wird also geteilt durch die Anzahl der Outlinks von W_i. Da mindestens $k \in O_i$, also für jeden Summanden immer $|O_i| > 0$ gilt, ist dieser Wert wohldefiniert.

Der PageRank-Wert basiert damit auf folgenden plausiblen Prinzipien (vgl.[Schö]):

- Je mehr Links auf eine Seite verweisen, desto bedeutender wird diese Seite.
  (*Je mehr Inlinks, desto mehr Summanden*)
- Je weniger ausgehende Links eine Seite enthält, desto bedeutender wird jeder einzelne Link.
  (*Je weniger Outlinks (je kleiner $|O_i|$), desto kleiner der jeweilige Summand*)
- Je bedeutender eine Seite ist, desto bedeutender sind die von ihr ausgehenden Links.
  (*Je größer x_i, desto größer der jeweilige Summand*)
- Je bedeutender die Links, die auf eine Seite verweisen, desto bedeutender die Seite.
  (*Je größer die einzelnen Summanden, desto größer die Summe*)

Beispiel:

Wir betrachten ein aus vier Seiten bestehendes Web mit der in Abbildung 1.3 graphisch veranschaulichten Link-Struktur:

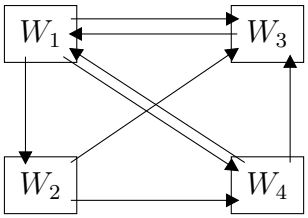


Abbildung 1.3: Beispiel-Webgraph für ein Web mit 4 Seiten

Nach der Formel (1.1) erhalten wir für diesen Webgraphen folgendes lineare Gleichungssystem:

$$\begin{array}{lllllllllllll}
x_1 & = & & & & \frac{x_3}{|O_3|} & + & \frac{x_4}{|O_4|} & = & & & \frac{x_3}{1} & + & \frac{x_4}{2} \\
x_2 & = & \frac{x_1}{|O_1|} & & & & & & = & \frac{x_1}{3} & & & & \\
x_3 & = & \frac{x_1}{|O_1|} & + & \frac{x_2}{|O_2|} & & + & \frac{x_4}{|O_4|} & = & \frac{x_1}{3} & + & \frac{x_2}{2} & + & \frac{x_4}{2} \\
x_4 & = & \frac{x_1}{|O_1|} & + & \frac{x_2}{|O_2|} & & & & = & \frac{x_1}{3} & + & \frac{x_2}{2} & &
\end{array}$$

Dieses Gleichungssystem kann mit Hilfe der Matrix-Schreibweise geschrieben werden als:

$$\underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix}}_{\text{Linkmatrix } A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}}_{x} \tag{1.2}$$

Hier stellt sich nun wieder die anfangs zurückgestellte Frage der Berechnung. Ist die Gleichung lösbar? Wenn ja, ist der Lösungsvektor eindeutig (bis auf Normierung) oder gibt es mehrere linear unabhängige Lösungsvektoren (und welchen sollte man dann als Rankingvektor verwenden)?

Ausgestattet mit den Grundkenntnissen aus der Linearen Algebra stellt man schnell fest, dass man es hier mit einem Eigenwertproblem zu tun hat.

Erinnerung:

Sei $A \in M_n(\mathbb{C})$.
Der Skalar $\lambda \, \epsilon \, \mathbb{C}$ heißt *Eigenwert* von A

$$:\Longleftrightarrow \exists \, 0 \neq x \in \mathbb{C}^n \quad \text{mit} \quad \boxed{\lambda x = Ax}.$$

Der Vektor x heißt dann *Eigenvektor* von A zum Eigenwert λ.

Mit anderen Worten: Die Gleichung (1.2) ist genau dann lösbar, wenn 1 ein Eigenwert der Linkmatrix A ist, d.h. ein Eigenvektor x zum Eigenwert 1 (= ein Fixvektor) existiert.

Der hier gewählte Webgraph und die dadurch bestimmte Linkmatrix ist unproblematisch: Aufgrund der Stochastizität der Linkmatrix A (siehe Kapitel 2) ist klar, dass 1 ein Eigenwert ist. Der zugehörige normierte Fixvektor x ist in diesem Beispiel eindeutig und lässt sich z.B. mit Programmen wie MatLab oder Octave schnell am Computer berechnen (gerundet auf 5 Dezimalen):

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \approx \begin{pmatrix} 0,38710 \\ 0,12903 \\ 0,29032 \\ 0,19355 \end{pmatrix}$$

Die Seite W_1 hätte im Beispiel-Web aus Abbildung 1.3 also den höchsten PageRank-Wert.

Bei einer standardmäßigen Eigenvektor-Berechnung einer Linkmatrix von der Größe, wie sie das WWW liefert, würden Programme wie Matlab oder Octave an dem zu bewältigenden Rechenaufwand scheitern. Google nutzt die sogenannte Potenzen-Methode. Hierbei macht auch die folgende Interpretation der Einträge der Linkmatrix Sinn:

Man stelle sich einen Internetuser vor, der nach dem Zufallsprinzip im Web surft. Wann immer er auf eine Seite stößt, folgt er zufällig einem der auf dieser Seite gesetzten Outlinks, um auf die nächste Seite zu gelangen, wo er nach dem gleichen Prinzip verfährt. Gelangt er also im Web von Abbildung 1.3 auf Seite W_1, so wird er mit einer Wahrscheinlichkeit von je $\frac{1}{3}$ als nächstes auf eine der Seiten W_2, W_3 oder W_4 klicken. Die Einträge a_{ij} der Linkmatrix A (siehe Gleichung (1.2)) geben also an, mit welcher Wahrscheinlichkeit der zufällige Surfer durch Anklicken eines Links von Seite W_j nach Seite W_i gelangt.

Nehmen wir an, der zufällige Surfer landet also nach dem ersten Link-Klick auf Seite W_2. Dort hat er zwei Möglichkeiten weiterzusurfen: Mit einem Link auf W_3 oder W_4 – bei jeweils 50%iger Wahrscheinlichkeit. Startet man also auf Seite W_1, so gelangt man mit Wahrscheinlichkeit $\frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$ mit zwei „Link-Schritten“ (über Seite W_2) zu Seite W_4. Will man in zwei Schritten von Seite W_1 auf Seite W_3 gelangen, so hat man zwei mögliche Wege: Entweder über die Seite W_2 oder über W_4 – jeweils mit Wahrscheinlichkeit $\frac{1}{3} \cdot \frac{1}{2}$. Die Wahrscheinlichkeit, dass der auf Seite W_1 startende zufällige Surfer nach zwei Schritten auf Seite W_3 landet, ist also $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$. Genau diese Transformations-Wahrscheinlichkeiten stehen in den Einträgen der Potenzen der Linkmatrix:

$$A^2 = \begin{pmatrix} \frac{1}{2} & \frac{3}{4} & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{3} & \frac{1}{6} \\ \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{6} \end{pmatrix}$$

Die Einträge $a_{ij}^{(2)}$ geben hier die Wahrscheinlichkeit an, mit welcher ein auf W_j startender zufälliger Surfer nach zwei Schritten auf Seite W_i landet.

Durch Induktion zeigt man leicht, dass auch für alle k die Matrixeinträge $a_{ij}^{(k)}$ von A^k die Wahrscheinlichkeit des zufälligen Surfens von W_j nach W_i in k-Schritten angeben (siehe z.B. [RobFo], Theorem 8.1). In diesem wie schon erwähnt besonders günstig gewählten Beispiel kann man feststellen, dass die Potenzen bzw. die Einträge der Matrix-Potenzen nach einigen Iterationen konvergieren (Einträge gerundet auf 5 Dezimalen):

$$\cdots,\ A^5 = \begin{pmatrix} 0.41667 & 0.43750 & 0.33333 & 0.37500 \\ 0.11111 & 0.12500 & 0.13889 & 0.15278 \\ 0.29167 & 0.27083 & 0.30556 & 0.27778 \\ 0.18056 & 0.16667 & 0.22222 & 0.19444 \end{pmatrix},$$

$$\cdots,\ A^{13} = \begin{pmatrix} 0.38725 & 0.38759 & 0.38662 & 0.38718 \\ 0.12887 & 0.12886 & 0.12924 & 0.12915 \\ 0.29038 & 0.29024 & 0.29036 & 0.29020 \\ 0.19350 & 0.19331 & 0.19377 & 0.19348 \end{pmatrix},$$

$$\cdots,\ A^{21} = \begin{pmatrix} 0.38710 & 0.38710 & 0.38709 & 0.38710 \\ 0.12903 & 0.12903 & 0.12903 & 0.12903 \\ 0.29032 & 0.29032 & 0.29032 & 0.29032 \\ 0.19355 & 0.19355 & 0.19355 & 0.19355 \end{pmatrix},$$

$$\cdots,\ A^{22} = \begin{pmatrix} 0.38710 & 0.38710 & 0.38710 & 0.38710 \\ 0.12903 & 0.12903 & 0.12903 & 0.12903 \\ 0.29032 & 0.29032 & 0.29032 & 0.29032 \\ 0.19355 & 0.19355 & 0.19355 & 0.19355 \end{pmatrix} = A^{23} = \cdots = A^{\infty}$$

Die Konvergenz gestaltet sich also hier so, dass in allen Spalten der Matrix A^{∞} genau der oben berechnete Fixvektor von A steht.

Dies lässt folgende Interpretation der Fixvektor-Einträge x_i zu: x_i ist die Wahrscheinlichkeit, dass man nach „ausreichend langem“ zufälligen Surfen auf der Seite W_i landet – egal von welcher Seite man gestartet ist.

Problem:

Von der unproblematischen Lösbarkeit in diesem besonders günstig gewählten Webgraph-Beispiel ist allerdings nicht auf eine grundsätzliche Lösbarkeit im Allgemeinen – d.h. für andere Webgraphen mit ungünstigeren Linkmatrizen – zu schließen. Der Wunsch nach Konvergenz für die Potenzen der Linkmatrix im Allgemeinen ist vielmehr utopisch.

Für ein Web mit n Seiten (WWW: $n =$ mehrere Milliarden) ergeben sich mit der PageRank-Formel (1.1) die folgende Gleichung und zwangsläufig die anschließenden Fragen:

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \overbrace{\underbrace{\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & & & \\ a_{31} & & \ddots & & \\ \vdots & & & \ddots & \\ a_{n1} & \cdots & \cdots & \cdots & a_{nn} \end{pmatrix}}_{wobei \begin{cases} a_{ij} = \frac{1}{|O_j|} \;\; falls \;\; j \in I_i \\ a_{ij} = \;\; 0 \quad\;\; sonst. \end{cases}}}^{\text{Linkmatrix } A} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} \tag{1.3}$$

Fragen:

1. Ist dieses Eigenwertproblem für die Linkmatrix A des riesengroßen World Wide Web lösbar? Wenn ja, ist der Fixvektor x bis auf Normierung eindeutig (d.h. es gibt keinen anderen linear unabhängigen Fixvektor)?

2. Konvergiert die Potenzen-Methode für die Linkmatrix, so dass der Fixvektor im Sinne der Wahrscheinlichkeits-Transformationen auch tatsächlich sinnvolle Rankingwerte enthält?

Ausblick:

In Kapitel 2 geht es nun darum, welche mathematischen Eigenschaften die Linkmatrix A und damit der zughörige Webgraph haben muss, damit die an Gleichung (1.3) anschließenden Fragen mit *ja* beantwortet werden können.

In Kapitel 3 wird sich dann zeigen, dass das WWW den Ansprüchen nicht genügt, und es wird dargestellt, wie Larry Page und Sergey Brin die Linkmatrix A sozusagen nach dem *Was-nicht-passt-wird-passend-gemacht*-Prinzip modifizieren, um letztendlich den PageRank-Fixvektor berechnen zu können. Darüber hinaus werden der Einfluss der entscheidenden Parameter und mögliche Veränderungen untersucht.

Kapitel 4 stellt eine alternative und weniger aufwendige mathematische Argumentation für das PageRank-Verfahren vor, um dann abschließend den Vorteil der in Kapitel 2 verwendeten spektraltheoretischen Argumentation herauszustellen und den Aufwand zu rechtfertigen.

2 Endlich-dimensionale Spektraltheorie

Wie schon am Ende von Kapitel 1 erwähnt, wird sich zeigen, dass die Linkmatrix A aus Gleichung (1.3), auf der das PageRank-Verfahren aufbaut, gewisse Anforderungen nicht erfüllt. Diese gewünschten aber nicht gegebenen (siehe Kapitel 3) mathematischen Eigenschaften werden zum Höhepunkt dieses Kapitels mit zwei Theoremen exakt formuliert.

2.1 Positive Matrizen

In diesem Abschnitt sollen Bedingungen gefunden werden, die hinreichend sind, um die erste der beiden an Gleichung (1.3) anschließenden Fragen mit *ja* zu beantworten.

Notation 2.1.1: Sei $x = (x_1, \dots, x_n)^T \in \mathbb{C}^n$

(a) Wir schreiben $x \geq 0$, falls $x_i \in \mathbb{R}$ und $\geq 0 \; \forall i$, und nennen x dann *positiv*.

(b) Wir schreiben $x \gg 0$, falls $x_i \in \mathbb{R}$ und $> 0 \; \forall i$, und nennen x dann *strikt positiv*.

(c) Mit *Betrag von* x ist der Vektor $|x| := (|x_1|, \dots, |x_n|)$ gemeint.

Notation 2.1.2: Sei $A = (a_{ij})_{i,j=1,\dots,n} \in M_n(\mathbb{C})$

(a) Wir schreiben $A \geq 0$, falls $a_{ij} \in \mathbb{R}$ und $\geq 0 \; \forall i, j$, und nennen A dann *positiv*.

(b) Wir schreiben $A \gg 0$, falls $a_{ij} \in \mathbb{R}$ und $> 0 \; \forall i, j$, und nennen A dann *strikt positiv*.

(c) Mit *Betrag von* A ist die Matrix $|A| := (|a_{ij}|)_{i,j=1,\dots,n}$ gemeint.

Definition 2.1.3:

(a) Sei $x = (x_1, \dots, x_n) \in \mathbb{C}^n$, dann heißt

$$\|x\|_1 := \sum_{i=1}^{n} |x_i| \qquad \textit{die 1-Norm auf } \mathbb{C}^n.$$

(b) Sei $A = (a_{ij})_{i,j=1,\dots,n} \in M_n(\mathbb{C})$, dann heißt

$$\|A\|_1 := max_{\|x\|_1=1} \|Ax\|_1 = max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}| = \textit{ maximale Spaltensumme von } A$$

die Matrixnorm auf $M_n(\mathbb{C})$.

Aufgrund der Normäquivalenz in endlich dimensionalen Räumen reicht es, sich hier auf diese Normen zu beschränken, die sich für die PageRank-Anwendung als passend erweisen. Der Nachweis der drei Normeigenschaften ist trivial.

Definition 2.1.4: Sei $\lambda \in \mathbb{C}$ Eigenwert von $A \in M_n(\mathbb{C})$, so ist

$$E_\lambda(A) := \{x |\ x \in \mathbb{C}^n \textit{ mit } Ax = \lambda x\} \text{ der } \textit{Eigenraum} \text{ von } A \text{ zum Eigenwert } \lambda,$$

und $dim\ E_\lambda(A)$ ist die geometrische Vielfachheit von λ.
Wenn 1 ein Eigenwert von A ist, so heißt $E_1(A) = Fix(A)$ der Fixraum von A.

Anmerkung: Wenn bisher die Rede war von der Suche nach einem eindeutigen Lösungsvektor der Gleichung (1.3), der sinnvolle Rankingwerte enthalten soll, dann ist also ein positiver (für die Rankingwerte x_i soll $x_i \geq 0\ \forall\ i$ gelten), auf 1 normierter Fixvektor gesucht, wobei $dim\ E_1(A) = dim\ Fix(A) = 1$ gilt (eindeutig).

Definition 2.1.5: Sei $A \in M_n(\mathbb{C})$. Dann ist

(a) die Menge aller Eigenwerte
$\sigma(A) := \{\lambda \in \mathbb{C}\ ;\ \lambda \text{ Eigenwert von } A\} = \{\lambda \in \mathbb{C}\ ;\ \exists\ 0 \neq x \in \mathbb{C}^n \text{ mit } Ax = \lambda x\}$
das Spektrum von A,

(b) der Betrag des betragsmäßig größten Eigenwerts
$r(A) := sup\{|\lambda| : \lambda \in \sigma(A)\}$
der Spektralradius von A.

Der folgende Satz ist von zentraler Bedeutung für die weitere Argumentation.

Satz 2.1.6 (Perron)[2]: *Sei A eine positive Matrix, dann ist r(A) ein Eigenwert von A mit zugehörigem positiven Eigenvektor.*

Beweis: Siehe z.B. [MacC], Perron's Theorem, oder [BNS], Theorem 1.6.4.

Korollar 2.1.7: *Sei $A \in M_n(\mathbb{C})$ und $\lambda \in \sigma(A)$, dann gilt $|\lambda| \leq \|A\|$*

Beweis: Nach Voraussetzung ist λ Eigenwert von A, d.h. $\exists\ 0 \neq x \in \mathbb{C}^n$ mit $Ax = \lambda x$, also auch $\|Ax\| = \|\lambda x\|$. Mit $\|Ax\| \leq \|A\|\ \|x\|$ und $\|\lambda x\| = |\lambda|\ \|x\|$ (Normeigenschaft) folgt $\|Ax\| = \|\lambda x\| = |\lambda|\ \|x\| \leq \|A\|\ \|x\|$. □

Definition 2.1.8: Eine Matrix $A \in M_n(\mathbb{C})$ nennt man *zeilen-* bzw. *spalten**sub**stochastisch*, wenn die Matrix positiv ist und die Summe der Einträge in jeder Zeile bzw. Spalte jeweils ≤ 1 ist. Man nennt solch eine positive Matrix A ***zeilen-*** bzw. ***spaltenstochastisch***, wenn die Zeilen- bzw. Spaltensumme sogar =1 ist für alle Zeilen bzw. Spalten.

Satz 2.1.9: *Sei $A \in M_n(\mathbb{C})$ spaltenstochastisch, dann gilt:*

(i) *1 ist Eigenwert von A,*
d.h. $\exists\ 0 \neq x \in \mathbb{C}^n$ mit $\lambda x = Ax$ für $\lambda = 1$.

(ii) *Es gibt keine betragsmäßig größeren Eigenwerte als 1,*
d.h. Spektralradius $= r(A) := sup\{|\lambda| : \lambda\ Eigenwert\} = 1$.

[2] Oskar Perron (1880-1975)

Beweis: (i) Sei $\mathbb{1} = (1, \ldots, 1) \in \mathbb{C}^n$. Es gilt $A\mathbb{1} = \mathbb{1}$ für A zeilenstochastisch, also 1 Eigenwert. Wegen $\sigma(A) = \sigma(A^T)$ folgt die Behauptung.
(ii) A spaltenstochastisch $\overset{(2.1.3)(b)}{\Longrightarrow} \|A\|_1 = 1 \overset{(2.1.7)}{\Longrightarrow} |\lambda| \leq 1 \; \forall \lambda \in \sigma(A) \overset{(2.1.5)(b)}{\Longrightarrow}$
$r(A) := sup\{|\lambda| : \lambda \in \sigma(A)\} = 1$. $\square$

Definition 2.1.10: Sei $A \in M_n(\mathbb{C})$. Ein Unterraum $Y \subset \mathbb{C}^n$ heiße *invariant unter* A, falls

$$Ay \in Y \; \forall \; y \in Y.$$

Definition 2.1.11: Eine Matrix $A \in M_n(\mathbb{C})$ heiße *reduzibel*, falls ein Unterraum

$$J_M := \{(\xi_1, \ldots, \xi_n) \; ; \; \xi_i = 0 \text{ für } i \in M\} \subset \mathbb{C}^n$$

für ein $\emptyset \neq M \subsetneq \{1, \ldots, n\}$ existiert, welcher invariant unter A ist. Falls A nicht reduzibel ist, so heiße A *irreduzibel*.

Oder: Die Matrix $A \in M_n(\mathbb{C})$ ist genau dann reduzibel, wenn es nach Umordnung der Basisvektoren von $\mathbb{C}^n$ ein $1 \leq k < n$ gibt, so dass

$$J_{M_k} := \{(\xi_1, \ldots, \xi_n) \; ; \; \xi_1 = \ldots = \xi_k = 0\}$$

invariant unter A ist.

Lemma 2.1.12: *Eine Matrix $A \in M_n(\mathbb{C})$ mit Matrixeinträgen a_{ij} ist genau dann irreduzibel, wenn zu jedem Indexpaar i,j mit $i \neq j$ eine Kette von Nicht-Nulleinträgen $a_{i,k_1}, a_{k_1,k_2}, a_{k_2,k_3}, ..., a_{k_{t-1},k_t}, a_{k_t,j}$ existiert.*

Beweis: Nach [Minc], Theorem 2.3, ist die Definition 2.1.11 einer irreduziblen Matrix A äquivalent mit der Aussage

$$\forall \; i, j \; \exists \; k \in \mathbb{Z}, \; \textit{so dass } a_{ij}^{(k)} > 0, \; \textit{wobei } a_{ij}^{(k)} \; \textit{die Einträge von } A^k \; \textit{sind.}$$

Interpretiert man die Einträge $a_{ij}^{(k)}$ wie in Abschnitt 1.4 als die Wahrscheinlichkeit, mit der ein zufälliger Surfer in k Schritten von Seite W_j zu Seite W_i gelangt, so bedeutet das, dass der zufällige Surfer, der auf einer beliebigen Webseite startet, irgendwann garantiert (d.h. mit Wahrscheinlichkeit > 0) auch zu jeder beliebigen anderen Webseite kommen kann. Es muss also für je zwei beliebige Webseiten einen Verbindungsweg über die Links geben. Mit der Erinnerung an die Tatsache, dass ein Linkmatrix-Eintrag a_{ij} genau dann $\neq 0$ ist, wenn ein Link von Seite W_j nach W_i existiert, wird schnell klar, dass die Existenz der Verbindungswege gleichbedeutend ist mit der Existenz der geforderten Ketten von Nicht-Nulleinträgen (vgl. auch [Schae], Kapitel III, 8). $\square$

Diese Charakterisierung von Irreduzibilität ist sehr nützlich, weil man beliebigen Webgraphen nun sehr schnell ansehen bzw. anmerken kann, ob ihre zugehörigen Linkmatrizen irreduzibel sind oder nicht: Eine Linkmatrix ist genau dann irreduzibel, wenn man von jeder beliebigen Webseite W_k auf jede andere Webseite W_l allein durch Klicken einer Kette von Links – also sozusagen, ohne die Hand von der Maus zu nehmen – gelangen kann.

Theorem 2.1.13 (Perron-Frobenius[3]): *Sei $A \in M_n(\mathbb{C})$ positiv und irreduzibel mit $r(A) = 1$. Dann ist 1 ein Eigenwert von A und der zugehörige Eigenraum ist eindimensional und wird von einem strikt positiven Vektor aufgespannt.*

Beweis: Nach Satz 2.1.6 wissen wir, dass $1 \in \sigma(A)$ ist und ein positiver Fixvektor z existiert. Also gilt $Az = z$ mit $0 \leq z = (\xi_1, \ldots, \xi_n)$. Angenommen, z sei nun nicht strikt positiv. Dann können wir nach Umordnung der Koordinaten annehmen, dass $\xi_i = 0$ für $i = 1, \ldots, k$ und $\xi_i > 0$ für $i = k+1, \ldots, n$. Also würde für jedes $y \in J_{M_k}$ (siehe 2.1.11) gelten, dass $|y| \leq c \cdot z$ für geeignete $c > 0$. Damit wäre

$$|Ty| \leq T|y| \leq cTz = c \cdot z,$$

was zeigt, dass $Ty \in J_{M_k}$, d.h. J_{M_k} ist T-invariant. Da T als irreduzibel vorausgesetzt wurde, ist dies ein Widerspruch. Also muss z strikt positiv sein.
Es ist nun noch zu zeigen, dass der Fixraum von A ein-dimensional ist: Angenommen, es existiere noch ein anderer linear unabhängiger Fixvektor $0 \neq y \in \mathbb{C}^n$ mit $Ay = y$. Da A eine positive, also eine reelle Matrix ist, folgt, dass Realteil und Imaginärteil von y Fixvektoren sind. Wir können also von $0 \neq y \in \mathbb{R}^n$ ausgehen. Da z strikt positiv ist, muss ein $c \in \mathbb{R}$ existieren, so dass

$$x := z - cy$$

positiv, aber nicht strikt positiv ist. Wie oben ist der den Null-Koordinaten von x zugehörige Unterraum J_M invariant unter A und muss daher 0 sein. Daraus folgt $z = cy$, was ein Widerspruch zur angenommenen linearen Unabhängigkeit von $\{z, y\}$ ist. □

Hinreichend für die Existenz eines eindeutigen (normierten) Fixvektors einer Matrix A sind also folgende Eigenschaften von A:

- *positiv*,
- *spaltenstochastisch*,
- *irreduzibel*.

Im nächsten Abschnitt wird sich zeigen, dass diese Liste noch um eine weitere Eigenschaft ergänzt werden muss, damit auch die Frage nach der Konvergenz der Potenzen-Methode (Frage 2 aus Kapitel 1) mit *ja* beantwortet werden kann und somit das PageRank-Verfahren tatsächlich funktioniert.

[3] Ferdinand Georg Frobenius (1849-1917)

2.2 Potenzen von Matrizen

Die Existenz eines eindeutigen, strikt positiven Fixvektors impliziert leider nicht die Anwendbarkeit der in Kapitel 1 vorgestellten Potenz-Methode zu seiner Berechnung.

Man betrachte beispielsweise folgenden Matrizen, die alle Anforderungen des Theorems 2.1.13 erfüllen:

$$B_n = \begin{pmatrix} 0 & 1 & 0 & \cdots & & 0 \\ & 0 & 1 & 0 & & \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & & \ddots & 1 & 0 \\ 0 & 0 & \cdots & & 0 & 1 \\ 1 & 0 & 0 & \cdots & & 0 \end{pmatrix}_{n\times n}$$

Eine solche „Permutationsmatrix“ $B_n \in M_n(\mathbb{C})$ ist positiv, spalten-(und zeilen-)stochastisch und irreduzibel. Der Spektralradius 1 ist Eigenwert und sein Eigenraum ist ein-dimensional und wird von einem strikt positiven Vektor (nämlich $\mathbb{1} = (1, \ldots, 1)$) aufgespannt. Wäre B_n eine von einem Web induzierte Linkmatrix, so würde sie dennoch den PageRank Ansprüchen nicht genügen: Die Potenzen-Methode würde hier scheitern. Betrachtet man die Potenzen von B_n für $n > 1$, so stellt man fest, dass sich diese periodisch verhalten mit

$$(B_n)^{kn+1} = B_n \; \forall \quad k \in \mathbb{N},$$

also nicht konvergieren. Die Eigenschaften positiv, stochastisch und irreduzibel allein reichen also noch nicht aus, um die PageRank-Ansprüche zu erfüllen.

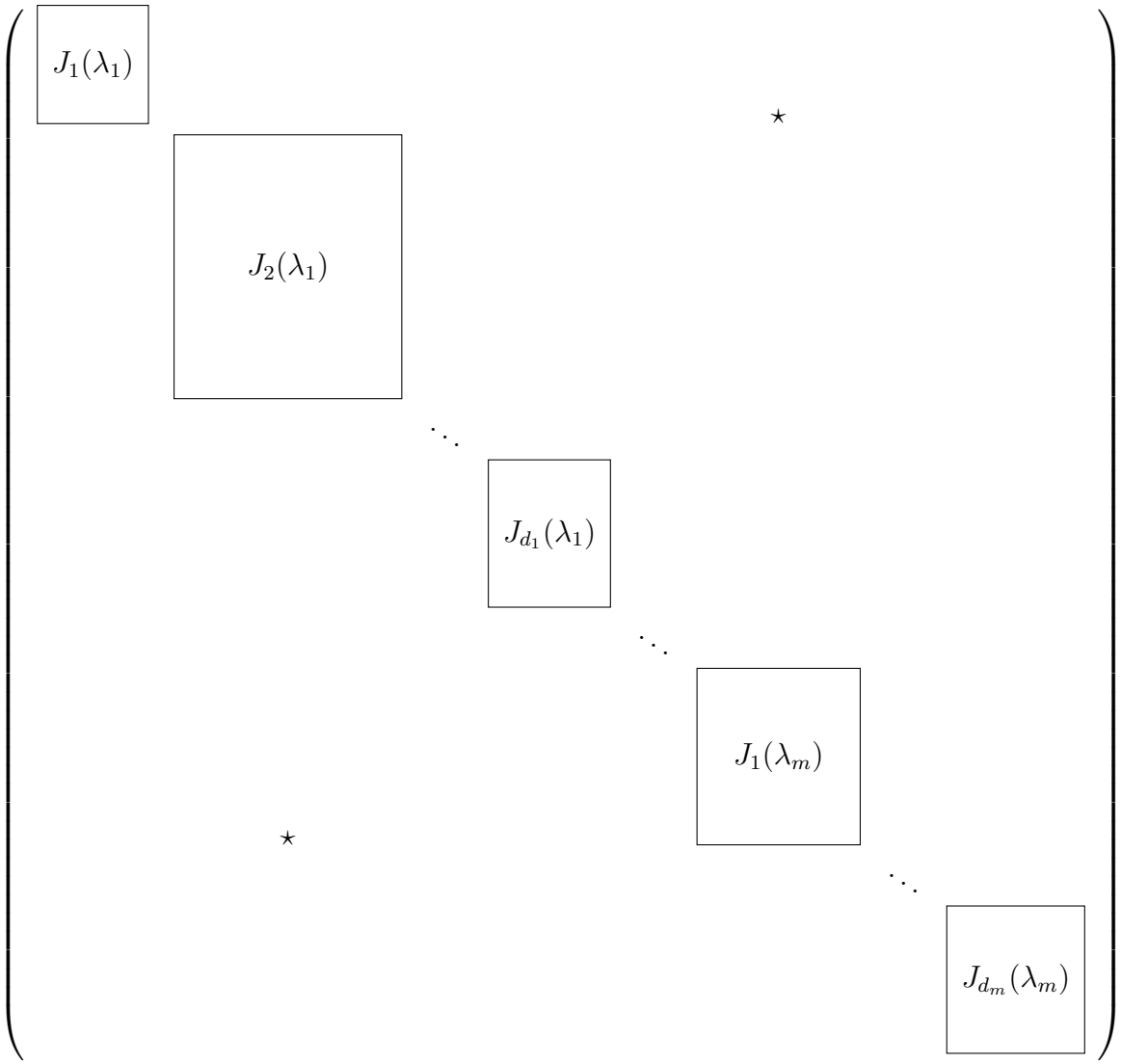
Um den PageRank-Anforderungskatalog vollständig zu formulieren zu können, ist es also nötig, sich näher mit Konvergenzkriterien für Potenzen positiver Matrizen auseinanderzusetzen. Dies soll in diesem Abschnitt geschehen.

Aus der Linearen Algebra wissen wir, dass „ähnliche“ Matrizen die gleichen elementaren Eigenschaften – wie beispielsweise das Konvergenzverhalten ihrer Potenzen – haben, weshalb man Matrizen gerne durch Ähnlichkeitstransformationen auf eine einfachere Form bringt. Eine solche Ähnlichkeitstransformation liefert die Jordan-Normalform einer Matrix, die im Folgenden für die Untersuchung von Konvergenzkriterien von großem Nutzen sein wird.

Erinnerung:

Es gibt zu jeder Matrix $A \in M_n(\mathbb{C})$ eine invertierbare Matrix S, so dass $S^{-1}AS$ Jordan-Normalform hat. Diese Jordan-Matrix $(S^{-1}AS)$ ist dann eine Matrix der gleichen Abbildung, die auch A induziert, nur bezüglich einer modifizierten Basis (vgl. [Fisch], Kapitel 4, Abschnitt 6).

Mit anderen Worten: Es gibt eine Basis $\{y_1, \ldots, y_n\}$ von $\mathbb{C}^n$, bezüglich derer die Matrix der von A definierten Abbildung auf $\mathbb{C}^n$ Jordan-Normalform $S^{-1}AS =: J =$



mit quadratischen Jordankästchen $J_k(\lambda_i)$ auf der Diagonalen (sonst ($\star$) nur Nullen) hat, wobei $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$ das Spektrum von A
und $d_i =$ *Anzahl der Jordankästchen zum Eigenwert* $\lambda_i = \dim E_{\lambda_i}$
ist. Jedes Jordankästchen besitzt die Form

$$J_k(\lambda_i) = \begin{pmatrix} \lambda_i & & & 0 \\ 1 & \lambda_i & & \\ & \ddots & \ddots & \\ 0 & & 1 & \lambda_i \end{pmatrix}$$

Mit $l(J_k(\lambda_i))$ sei die Länge des Jordankästchens $J_k(\lambda_i)$ bezeichnet, und sei $l_i = \sum_j l(J_k(\lambda_i))$ die Gesamtlänge aller zum Eigenwert λ_i gehörenden Jordankästchen zusammen = die Länge des gesamten zu λ_i gehörenden quadratischen „Jordanblocks“.

Mit dieser Notation gilt Folgendes für die Hauptdiagonale der Matrix J: Von der ersten Spalte bis zur Spalte $j_1 = l_1$ steht der Eigenwert λ_1 auf der Hauptdiagonalen. Von Spalte $j_1 + 1$ bis Spalte $j_2 = l_1 + l_2$ ist λ_2 eingetragen. Der Eigenwert λ_i steht von Spalte $j_{i-1} + 1$ bis Spalte $j_i = l_1 + l_2 + \ldots + l_i$ in der Hauptdiagonalen. Ferner setze aus Notationsgründen $j_0 = 0$.

Im Folgenden soll dieses Vorwissen über die Jordan-Normalform weiter vertieft werden.

Definition 2.2.1: Ein lineare Abbildung P heißt Projektion, wenn sie idempotent ist, d.h. wenn $P^2 = P$ gilt.

Satz 2.2.2: *Wenn $P \in M_n(\mathbb{C})$ eine Projektion ist, dann gilt $\mathbb{C}^n = \operatorname{im} P \oplus \ker P$.*

Beweis: Sei $x \in \mathbb{C}^n$ und setze $u := Px$ und $w := x - u$. Weil P linear und idempotent ist, gilt dann $Pw = Px - Pu = Px - P^2x = 0$. Folglich ist $w \in \ker P$. Es ist nun auch klar, dass $Pu = u$ für alle $u \in \operatorname{im} P$ gilt. Die Summe ist direkt: $x \in \operatorname{im} P \cap \ker P$ liefert $x = Px = 0$. $\square$

Satz 2.2.3: *Für jede direkte Zerlegung $\mathbb{C}^n = U \oplus W$ existiert eine eindeutig bestimmte Projektion P mit* $\operatorname{im} P = U$ *und* $\ker P = W$.

Beweis: Für $x = u + w$ mit $u \in U$ und $w \in W$ setze $Px := u$. Offenbar ist auch $P^2x = u$, also P eine Projektion. Die Eindeutigkeit ist klar.

Nun kann die folgende wichtige Klasse von Projektionen eingeführt werden:

Definition 2.2.4: Sei $A \in M_n(\mathbb{C})$, $\sigma(A) = \{\lambda_1, \ldots, \lambda_m\}$ und sei $\{y_1, \ldots, y_n\}$ Basis, bezüglich derer die Matrix der von A definierten Abbildung Jordan-Normalform J hat (wie zuvor). Die Projektion auf $< y_1, \ldots, y_{j_1} >$ mit Kern $< y_{j_i+1}, \ldots, y_n >$ heißt Spektralprojektion von A zum Eigenwert λ_1 und wird mit P_1 bezeichnet.

Die Matrix von P_1 bezüglich der Basis $\{y_1, \ldots, y_n\}$ hat die Form

$$\begin{array}{c} \text{Spalte } j_1 \\ \downarrow \\ \begin{pmatrix} 1 & & \vdots & & \\ & \ddots & \vdots & \star & \\ & & 1 & & \\ & & & 0 & \\ & \star & & & \ddots \\ & & & & & 0 \end{pmatrix} \end{array}$$

mit ausschließlichen Null-Einträgen in allen Nebendiagonalen ($\star$).

Ebenso bekommt man für die Eigenwerte $\lambda_2, \ldots, \lambda_m$ die Spektralprojektionen $P_2, \ldots, P_m$, wobei P_i die Projektion auf $< y_{j_{i-1}+1}, \ldots, y_{j_i} >$ mit Kern $< \bigcup_{k \notin \{j_{i-1}+1, \ldots, j_i\}} y_k >$ ist.

Die Matrix von P_i hat bezüglich der Basis $\{y_1, \ldots, y_n\}$ die Form

$$\begin{pmatrix}
 & & j_{i-1}+1 & & j_i & & \\
0 & & \vdots & & \vdots & & \\
 & \ddots & \vdots & & \vdots & \star & \\
 & & 0 \;\vdots & & \vdots & & \\
 & & 1 & & \vdots & & \\
 & & & \ddots & \vdots & & \\
 & & & & 1 & & \\
 & & & & & 0 & \\
 & \star & & & & & \ddots \\
 & & & & & & 0
\end{pmatrix}.$$

Lemma 2.2.5: Sei $A \in M_n(\mathbb{C})$. Für die durch die Eigenwerte eindeutig bestimmten Spektralprojektionen $P_1, \ldots, P_m$ gilt

$$P_1 + P_2 + \ldots + P_m = id \qquad \text{und} \qquad P_i P_j = 0 \text{ } \textit{für } i \neq j.$$

Mit der Definition der Spektralprojektionen und der Feststellung aus Lemma 2.2.5 lässt sich die Jordan-Normalform von A folgendermaßen beschreiben:

$$S^{-1}AS =: J = JI = J\left(\sum_{i=1}^{m} P_i\right) = \sum_{i=1}^{m} JP_i \quad . \tag{2.1}$$

Die Matrix von JP_i sieht dabei bezüglich der Basis $\{y_1, \ldots, y_n\}$ folgendermaßen aus:

$$JP_i = \begin{pmatrix}
0 & & & & & & & \\
 & \ddots & & & & & & \\
 & & 0 & & & & \star & \\
 & & & \boxed{J_1(\lambda_i)} & & & & \\
 & & & & \ddots & & & \\
 & & & & & \boxed{J_{d_i}(\lambda_i)} & & \\
 & & \star & & & & & \\
 & & & & & & 0 & \\
 & & & & & & & \ddots \\
 & & & & & & & 0
\end{pmatrix}$$

Durch weitere Umformung der in Gleichung (2.1) erhaltenen Summendarstellung ergibt sich

$$J = \sum_{i=1}^{m} JP_i = \sum_{i=1}^{m} (\lambda_i + (J - \lambda_i))P_i \quad . \tag{2.2}$$

Doch nun zurück zum Verhalten der Potenzen einer Matrix.

Ziel des Abschnitts war es, Konvergenzkriterien für Matrix-Potenzen zu erhalten. Die bis hierher geleistete Vorarbeit mit der Erinnerung an die Jordan-Normalform und ihre Beschreibung mit Hilfe der eingeführten Spektralprojektionen dienen nun dazu, das Potenz-Verhalten von A analysieren zu können.

Für die Potenzen J^k gilt mit (2.2):

$$\begin{aligned} J^k &= \left(\sum_{i=1}^{m} (\lambda_i + (J - \lambda_i))P_i \right)^k \\ &\overset{P_iP_j=0, i\neq j}{=} \sum_{i=1}^{m} \left((\lambda_i + (J - \lambda_i))P_i \right)^k . \end{aligned}$$

Nach Anwendung der Binomischen Formel erhält man schließlich

$$J^k = \sum_{i=1}^{m} \left\{ \sum_{\nu=0}^{k} \binom{k}{\nu} \lambda_i^{k-\nu} (J - \lambda_i)^\nu P_i \right\} . \tag{2.3}$$

Man betrachte nun die Matrix von $(J - \lambda_i)P_i$ bezüglich $\{y_1, \ldots, y_n\}$:

$$(J - \lambda_i)P_i = \begin{pmatrix} \ddots & & & j_{i-1}+1 & & & & j_i & \\ & \ddots & & \vdots & & & & \vdots & \star \\ & & \ddots & \vdots & & & & \vdots & \\ & & & \vdots & & & & \vdots & \\ & & & 0 & & & & \vdots & \\ & & & \epsilon_{j_{i-1}+1} & \ddots & & & \vdots & \\ & & & & \ddots & \ddots & & \vdots & \\ & & & & & \ddots & \ddots & \vdots & \\ & & & & & & \ddots & \vdots & \\ & & & & & & \epsilon_{j_1} & 0 & \\ & & & & & & & & \ddots \\ \star & & & & & & & & \ddots \end{pmatrix}$$

mit $\epsilon_j = 0$ oder 1 (je nach Anzahl und Größe der einzelnen Blöcke zum Eigenwert λ_i).

Es ist leicht festzustellen, dass die Einträge ϵ_j mit jeder höheren Potenz von $(J - \lambda_i)P_i$ eine Nebendiagonale nach unten wandern, bis sie irgendwann komplett aus dem Block zu

λ_i verschwunden sind und nur noch Nullen in der Matrix-Potenz übrig sind.

Mit anderen Worten: $(J-\lambda_i)P_i$ ist nilpotent. Wir bezeichnen die Nilpotenzordnung von $(J-\lambda_i)P_i$ – d.h. die kleinste natürliche Zahl ν, für die $((J-\lambda_i)P_i)^\nu = (J-\lambda_i)^\nu P_i = 0$ gilt – mit ν_i. Damit ist klar, dass

$$\nu_i = \max_k l(J_k(\lambda_i)) = \textit{ Länge des größten zu } \lambda_i \textit{ gehörenden Jordankästchens}$$

ist.

Sind beispielsweise alle Jordan-Kästchen zu λ_i 1×1-Matrizen, also somit

$$\max_k l(J_k(\lambda_i)) = \nu_i = 1,$$

dann ist schon $(J-\lambda_i)P_i = 0$. In diesem Fall sagt man, dass λ_i ein *einfacher Eigenwert* ist.

Aus dem Wissen über die Nilpotenzordnung von $(J-\lambda_i)^\nu P_i$ folgt, dass die Summe $\sum_{\nu=0}^{k}(J-\lambda_i)^\nu P_i$ aus (2.3) ab dem ν_i-ten Summanden nur noch Nullen enthält. Wir müssen also nur bis zum (ν_i-1)-ten Summanden aufsummieren und erhalten damit für die Potenzen J^k die Darstellung

$$J^k = \sum_{i=1}^{m}\sum_{\nu=0}^{\nu_i-1}\binom{n}{\nu}\lambda_i^{n-\nu}(J-\lambda_i)^\nu P_i. \tag{2.4}$$

Lemma 2.2.6: *Die Menge*

$$\mathcal{B}_J := \{(J-\lambda_i)^\nu P_i \;;\; i=1,\dots,m \;;\; \nu = 0,\dots,\nu_i-1\}$$

ist linear unabhängig in $M_n(\mathbb{C})$.

Beweis: Betrachtet man $(J-\lambda_i)^\nu P_i$ und die jeweils ν-ten Potenzen (siehe vorangegangene Diskussion) für $i=1,\dots,m$ und $\nu = 0,\dots,\nu_i-1$, so ist klar, dass je zwei verschiedenen Elemente von $\mathcal{B}_J$ ihre Eins-Einträge auf verschiedenen Positionen haben (vgl. [Wint], S.10). □

Nach dem Basisergänzungssatz kann nun eine Basis $\mathcal{B}$ von $M_n(\mathbb{C})$ so gewählt werden, dass sie die linear unabhängige Menge $\mathcal{B}_J$ enthält. Die Darstellung (2.4) von J^k zeigt, dass

$$\left\{\binom{k}{\nu}\lambda_i^{k-\nu} \;;\; i=1,\dots,m \;;\; \nu=0,\dots,\nu_i-1\right\}$$

die Menge aller Nicht-Null-Einträge von J^k ist (bezüglich $\mathcal{B}$).

Da unter Konvergenz von Matrizen die Konvergenz in jedem Eintrag verstanden wird, ist für die Existenz des Limes

$$\lim_{k\to\infty} J^k = \lim_{k\to\infty}\left(\sum_{i=1}^{m}\sum_{\nu=0}^{\nu_i-1}\binom{k}{\nu}\lambda_i^{k-\nu}(J-\lambda_i)^\nu P_i\right)$$

also nur das Verhalten der Folgen

$$z_{\lambda_i,\nu}(k) := \binom{k}{\nu}\lambda_i^{k-\nu}$$

für alle $\lambda_i \in \sigma(A),\ \nu = 0, \ldots, \nu_i - 1$ entscheidend.

Dieses Verhalten ist aber einfach zu verstehen und hauptsächlich vom Betrag von λ_i abhängig:

- Falls $|\lambda_i| < 1$, dann $z_{\lambda_i,\nu}(k) \overset{k\to\infty}{\longrightarrow} 0$ für alle ν.
- Falls $|\lambda_i| > 1$, dann $|z_{\lambda_i,\nu}(k)| \overset{k\to\infty}{\longrightarrow} \infty$ für alle ν.
- Falls $|\lambda_i| = 1$ und $\nu_i - 1 = 0$ (also $\nu_i = 1$), dann $z_{\lambda_i,\nu}(k) = \lambda_i^k$.
- Falls $|\lambda_i| = 1$ und $\nu_i - 1 > 0$ (also $\nu_i > 1$), dann $|z_{\lambda_i,\nu}(k)| \overset{k\to\infty}{\longrightarrow} \infty$.

Damit gelten folgende Feststellungen für die Matrix J (stets $|\lambda_i^\star| = r(J)$):

(a) Ist der Spektralradius $r(J) > 1$, so sind die Potenzen von J unbeschränkt (weil mindestens der Eintrag $(\lambda_i^\star)^k$ unbeschränkt).

(b) Ist der Spektralradius $r(J) < 1$, so gilt $J^k \overset{k\to\infty}{\longrightarrow} 0$ (weil $|\lambda_i| < 1$ für alle λ_i).

(c) Ist der Spektralradius $r(J) = 1$, so gilt Folgendes:

- Falls $\nu_i > 1$ für ein λ_i mit $|\lambda_i| = 1$, so ist J^k unbeschränkt.
- Falls $\exists\ \lambda_i \neq 1$ mit $|\lambda_i| = 1$, so ist J^k nicht konvergent.
- $0 \neq \lim\limits_{k\to\infty} J^k$ existiert
  $$\iff \begin{cases} \lambda_i^\star = 1,\ \nu_i^\star = 1 \text{ (d.h. 1 ist einfacher Eigenwert)} \\ \text{und} \\ |\lambda_i| < 1 \text{ für alle } \lambda_i \neq 1 \text{ (d.h. 1 ist „dominanter“ Eigenwert)} \end{cases}$$

Um das Ziel der Untersuchungen nicht aus den Augen zu verlieren, sei daran erinnert, dass hier ohnehin nur der Fall $r(J) = 1$ interessiert. Gesucht wird nach Konvergenzkriterien für Potenzen einer Matrix, von der schon klar ist, dass sie *positiv, spaltenstochastisch* (folglich ist der Spektralradius =1, siehe Satz 2.1.9) und *irreduzibel* sein muss (siehe Theorem 2.1.13).

Um den PageRank-Anforderungskatalog zu vervollständigen, muss die Liste also nur noch um die Forderung nach Einfachheit und Dominanz des Eigenwerts 1 ergänzt werden. Das folgende Theorem wird zeigen, dass die Einfachheit vom Eigenwert 1 ohnehin schon impliziert ist. Die eigentliche Anforderungs-Erweiterung besteht also nur in der Forderung nach Dominanz von 1.

Theorem 2.2.7a: *Sei $A \in M_n(\mathbb{C})$ positiv und irreduzibel und sei $\lambda_1 = 1$ dominanter Eigenwert von A. Sei $J = S^{-1}AS \in M_n(\mathbb{C})$ die Jordan-Normalform, die die von A definierte Abbildung bezüglich einer modifizierten Basis hat, und sei P_1 die Spektralprojektion zum Eigenwert 1. Dann gilt*

$$\lim_{k\to\infty} J^k = \lim_{k\to\infty} \left(\sum_{i=1}^{m} \sum_{\nu=0}^{\nu_i-1} \binom{k}{\nu} \lambda_i^{k-\nu} (J-\lambda_i)^\nu P_i \right) = P_1.$$

Beweis: Zunächst muss noch gezeigt werden, dass 1 ein einfacher Eigenwert ist.
Nach Theorem 2.1.13 existiert ein strikt positiver Vektor $z = (\xi_1, \ldots, \xi_n)$ mit

$$Az = z.$$

Definiere $D := \mathrm{diag}(\xi_1, \ldots, \xi_n)$ und

$$T := D^{-1}AD.$$

Dann gilt $0 \leq T$ und $Te = e = (1, \ldots, 1)$, folglich ist $\|T\| = 1$. Daraus folgt, dass $\|T^k\| \leq 1$ und

$$\|A^k\| \leq \|DT^kD^{-1}\| \leq \|D\| \cdot \|D^{-1}\|$$

für alle $k \in \mathbb{N}$. Die Potenzen A^k und damit auch $J^k = (S^{-1}AS)^k = S^{-1}A^kS$ sind also für $k \to \infty$ beschränkt. Die Formel (2.4) liefert nach Multiplikation mit der zum Eigenwert 1 gehörenden Projektion P_1 die Gleichung

$$J^kP_1 = \sum_{\nu=0}^{\nu_1-1} \binom{k}{\nu} (J-1)^\nu P_1 \quad .$$

Diese Summe ist allerdings nur dann beschränkt für $k \to \infty$, wenn $\nu_1 = 1$ ist.

Wegen der vorausgesetzten Dominanz von $\lambda_1 = 1$ gilt dann

$$z_{\lambda_i,\nu}(k) := \binom{k}{\nu} \lambda_i^{k-\nu} \to 0 \;\forall\; \lambda_i \neq \lambda_1$$

und damit

$$\lim_{k\to\infty} J^k = \lim_{k\to\infty} \left(\sum_{\nu=0}^{\nu_i-1} \binom{k}{\nu} \lambda_1^{k-\nu} (J-\lambda_1)^\nu P_1 \right).$$

Mit $\nu_1 = 1$ folgt

$$\lim_{k\to\infty} J^k = \lim_{k\to\infty} \left(\binom{k}{0} \lambda_1^{k-0} (J-\lambda_1)^0 P_1 \right) = P_1. \quad \square$$

Theorem 2.2.7b: *Voraussetzungen wie in Theorem 2.2.7a. Sei A zusätzlich noch spaltenstochastisch. Dann gilt*

$$\lim_{k\to\infty} A^k = B := SP_1S^{-1},$$

wobei $\operatorname{im} B = Fix(A)$ *ist und alle Spaltenvektoren von B identisch sind mit dem eindeutigen, strikt positiven Fixvektor von A, der Koordinatensumme 1 hat.*

Beweis: Es gilt

$$\lim_{k\to\infty} \|A^k - SP_1S^{-1}\| = \lim_{k\to\infty} \|S(S^{-1}A^kS - P_1)S^{-1}\| \le \|S\| \, \| \underbrace{S^{-1}A^kS - P_1}_{\overset{k\to\infty}{\longrightarrow} 0 \,(mit\ a)} \| \, \|S^{-1}\|.$$

$\Rightarrow \lim_{k\to\infty} A^k = SP_1S^{-1} = B \quad .$

Auch B ist eine Projektion, denn $SP_1 \underbrace{S^{-1} \cdot S}_{I} P_1S^{-1} = SP_1S^{-1}$.

Für $z \in \operatorname{im} B$, etwa $z = By$ mit $y \in \mathbb{C}^n$, gilt folglich

$$Bz = B^2y = By = z.$$

Damit ist

$$Az = ABz = A\left\{\left(\lim_{k\to\infty} A^k\right) z\right\} = \left(\lim_{k\to\infty} A^{k+1}\right) z = Bz = z,$$

d.h. $z \in Fix(A)$.

Seien $e_k \in \mathbb{C}^n$ $(k = 1, 2, \ldots, n)$ die kanonischen Basisvektoren von $\mathbb{C}^n$ (k-te Koordinate =1, sonst nur Nullen). Dann gilt für die Vektoren Be_k (= Spaltenvektoren von B)

$$Be_k \in Fix(A) \ \ \forall\, k = 1, 2, \ldots, n$$

mit $Be_k \ge 0$ ($B \ge 0$, da $B = \lim_{k\to\infty} A^k$ & $A^k \ge 0$) und sogar $Be_k \gg 0$ (Theorem 2.1.13).

Da der Fixraum von A nach Theorem 2.1.13 eindimensional ist, müssen alle Spaltenvektoren von B linear abhängig sein. Es bleibt also noch zu zeigen, dass sie sogar alle identisch sind. Hierbei geht nun die Stochastizität von A ein:

$$\begin{aligned}
\langle Be_k, \mathbb{1}\rangle &= \langle e_k, B^T\mathbb{1}\rangle \\
&= \left\langle e_k, \underbrace{\lim_{k\to\infty} (A^T)^k \cdot \mathbb{1}}_{\underbrace{(A^T \ldots A^T \underbrace{A^T)\mathbb{1}}_{=\mathbb{1}}}_{=\mathbb{1}}} \right\rangle \\
&= \langle e_k, \mathbb{1}\rangle = 1 \quad \forall\, k = 1, 2, \ldots, n. \qquad \square
\end{aligned}$$

Definition 2.2.8: Sei $A \in M_n(\mathbb{C})$ positiv und irreduzibel und sei der Spektralradius $r(A)$ ein dominanter Eigenwert, dann nennt man A *primitiv*.

Der folgende Satz, der auch auf Oskar Perron zurückgeht, stellt eine elegante Anwendung von Satz 2.1.6 dar und wird später eine wichtige Rolle spielen.

Satz 2.2.9 (vgl. z.B.[LaMey], S.174, oder [Minc], Kapitel 3)**:**
Falls $A \in M_n(\mathbb{C})$ *strikt positiv ist, dann ist* A *primitiv.*

Beweis: Da die Matrix A strikt positiv ist, ist sie selbstverständlich auch irreduzibel (da alle Einträge > 0 sind, existiert natürlich auch für jedes Indexpaar eine Kette von Nicht-Nulleinträgen im Sinne von Lemma 2.1.12). Es ist also nur noch zu zeigen, dass der Spektralradius $r(A)$ ein dominanter Eigenwert ist.

Wegen $A = (a_{ij})_{n\times n} \gg 0$ gilt insbesondere $\inf_i a_{ii} > 0$ und nach Satz 2.1.6 $r(A) > 0$.
Es gilt

$$\sigma(A - \rho I) = \sigma(A) - \rho = \{\lambda - \rho \mid \lambda \in \sigma(A)\} \ \forall\, \rho.$$

Anschaulich heißt das, dass alle Eigenwerte von $A - \rho I$ innerhalb des um ρ nach links verschobenen Spektralkreises von A liegen müssen, also innerhalb des grünen Kreises von Abbildung 2.1.

Sei $\rho = \inf_i a_{ii}$. Dann gilt $A - \rho I \geq 0$.
Mit Satz 2.1.6 folgt, dass $r(A - \rho I)$ ein Eigenwert von $A - \rho I$ ist. Also gilt $r(A - \rho I) = r(A) - \rho$, was wiederum anschaulich bedeutet, dass die Eigenwerte von $A - \rho I$ sogar ausschließlich innerhalb des blauen Kreises von Abbildung 2.1 liegen müssen. Damit ist klar, dass $\sigma(A) = \sigma((A - \rho I) + \rho)$ gilt, also die Eigenwerte von A tatsächlich innerhalb des roten Kreises von Abbildung 2.1 liegen. Folglich ist $r(A)$ der einzige Eigenwert auf dem Spektralkreis von A, also dominant. □

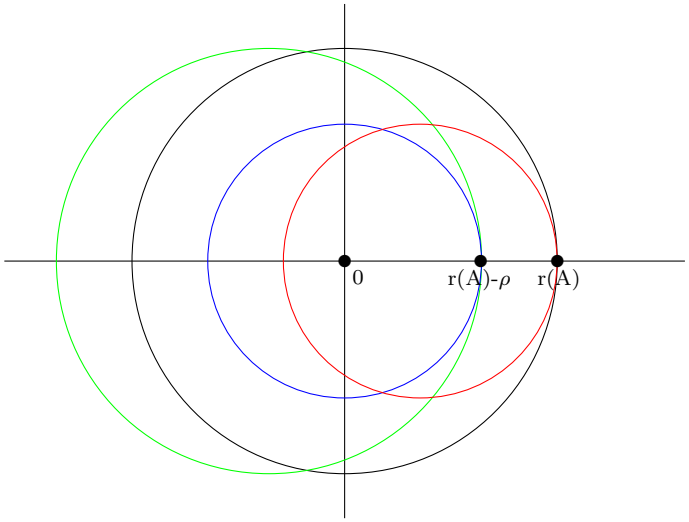


Abbildung 2.1: Zur Lokalisierung von $\sigma(A)$ und $\sigma(A - \rho I)$

3 Die Google-Matrix

Kapitel 1 endete mit der Frage, welche Eigenschaften die Linkmatrix haben muss, damit der PageRank-Ansatz funktioniert, d.h. die Gleichung (1.3) eindeutig lösbar ist und die Potenzen der Linkmatrix konvergieren.

Die Ergebnisse von Kapitel 2 haben folgende Wunschliste hinreichender Bedingungen geliefert: Die Linkmatrix A muss

- *positiv*,
- *spaltenstochastisch*

und

- *irreduzibel*

sein und

- *der Spektralradius 1 muss dominant sein.*

Sicher erfüllt ist für die Linkmatrix allerdings nur die Positivität (alle Einträge entweder 0 oder $\frac{1}{|O_j|}$). Der nächste Abschnitt wird zeigen, dass die restlichen Ansprüche utopisch sind. Anschließend wird es darum gehen, wie die den Ansprüchen nicht genügende Linkmatrix des WWW durch clevere Modifizierungen in eine doch PageRank-fähige Matrix – die Google-Matrix G – umgewandelt wird.

3.1 Problematische Webstrukturen

3.1.1 Dangling Nodes (Webseiten ohne Outlinks)

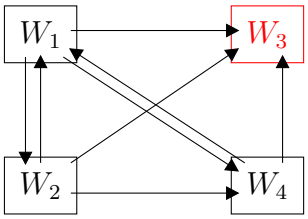


Abbildung 3.1: Beispiel-Webgraph für ein Web mit Dangling Node

Der Webgraph in Abbildung 3.1 zeigt eine Web-Konstellation, die im WWW unzählige Male vorkommt: Viele Webseiten besitzen keine Outlinks. Man denke hier nur beispielsweise an die zahlreichen neu ins Netz gestellten Seiten, von denen noch keine Links gesetzt wurden, an PDF-Dokumente, auf die man zwar durch eine Verlinkung gelangt, von denen aber kein Link mehr wegführt, oder an Web-Eigenbrötler, die einfach nicht an einer

Verlinkung mit anderen Seiten interessiert sind. Um im Bild von Kapitel 1 zu bleiben: In der demokratischen WWW-Wahl gibt es also zahlreiche Webautoren, die nicht von ihrem Wahlrecht Gebrauch machen.

Genauso wie eine Gesellschaft es nicht gutheißt, wenn es viele Nichtwähler gibt, so wenig erfreulich sind die Dangling Nodes für die PageRank-Ansprüche. Denn die zu den Dangling Nodes gehörenden Spalten der Linkmatrix enthalten ausschließlich Null-Einträge. Die Linkmatrix des Webgraphen aus Abbildung 3.1 wäre beispielsweise

$$A_{[3.1]} = \begin{pmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \end{pmatrix}.$$

Mit anderen Worten: Die Linkmatrizen von Webs mit Dangling Nodes sind nicht spaltenstochastisch, sondern nur spaltensubstochastisch. Also ist die Linkmatrix des WWW definitiv nicht spaltenstochastisch.

3.1.2 Nicht stark zusammenhängender Webgraph

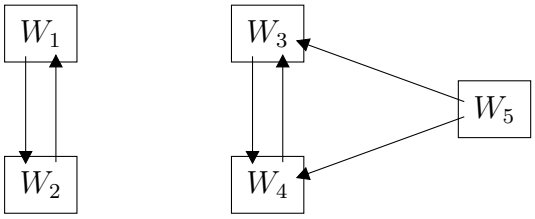


Abbildung 3.2: Beispiel-Webgraph für ein nicht stark zusammenhängendes Web

Der Begriff des starken Zusammenhangs stammt aus der Graphentheorie und ist gleichbedeutend mit Irreduzibilität der Inzidenzmatrix. Auch ohne graphentheoretisches Wissen wird jedoch schnell klar, dass der Webgraph aus Abbildung 3.2 eine nicht irreduzible Linkmatrix mit dementsprechend nicht eindimensionalem Eigenraum zur 1 erzeugt:

$$A_{[3.2]} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}; \quad E_1(A_{[3.2]}) = \left\{ \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \\ 0 \end{pmatrix} s, \begin{pmatrix} 0 \\ 0 \\ \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{pmatrix} t \,;\, s,t \in \mathbb{R} \right\}$$

Es ist offensichtlich, dass hier nicht zu jedem Indexpaar i,j mit $i \neq j$ eine Kette von Nicht-Nulleinträgen $a_{i,k_1}, a_{k_1,k_2}, a_{k_2,k_3}, ..., a_{k_{t-1},k_t}, a_{k_t,j}$ existiert. Mit anderen Worten: Man kann nicht von jeder Webseite nur durch Klicken von Links auf jede andere Webseite gelangen. Das wiederum heißt nach Lemma 2.1.12, dass die Matrix nicht irreduzibel ist.

Auch diese Konstellation ist im WWW natürlich nicht die Ausnahme, sondern vielmehr die Regel. Es gibt Millionen von Unterwebs, die nicht miteinander verbunden sind. Die Linkmatrix des WWW ist also ohne Zweifel nicht irreduzibel.

3.2 Modifizierung der Linkmatrix A

Die vom WWW durch die PageRank-Formel (1.1) erzeugte Linkmatrix A ist also zwar positiv, aber weder stochastisch noch irreduzibel. Somit sind die Voraussetzungen aus Theorem 2.1.13 für die Existenz eines strikt positiven, eindeutigen Fixvektors nicht gegeben.

Die Linkmatrix A genügt also nicht den Ansprüchen von Larry Page und Sergey Brin und dennoch funktioniert PageRank – wie kann das sein? Eigentlich ganz einfach: Was nicht passt, wird passend gemacht! Die Linkmatrix wird einfach – ohne dabei den Charakter der ursprünglichen Linkmatrix als Reflektion der demokratischen Web-Abstimmung vollständig zu zerstören – so modifiziert, dass das PageRank-Verfahren problemlos funktioniert. Dies soll in den folgenden Abschnitten dargestellt werden.
Die Modifikationen werden anhand des Beispiel-Webgraphen in Abbildung 3.3, der die beiden schon vorgestellten problematischen Webstrukturen enthält, anschaulich gemacht:

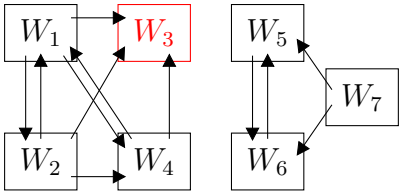


Abbildung 3.3: Nicht stark zusammenhängender Webgraph mit Dangling Node

3.2.1 Behebung des Dangling Node-Problems

Im ersten Schritt werden alle diejenigen Spalten der $n \times n$-Linkmatrix, die wegen der Dangling Nodes nur Nullen enthalten, durch einen Spaltenvektor ersetzt, der $\frac{1}{n}$ in jeder Koordinate stehen hat.
Für die Linkmatrix $A_{[3.3]}$ des Webgraphen aus Abbildung 3.3 sieht dies folgendermaßen aus:

$$\underbrace{\begin{pmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}}_{=: \textit{Linkmatrix } A_{[3.3]}} + \underbrace{\begin{pmatrix} 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}}_{=: \textit{Dangling Node}-\textit{Matrix } D_{[3.3]}} = \underbrace{\begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{7} & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}}_{=: \textit{Stochastische Matrix } S_{[3.3]}}$$

Allgemein gilt für die Matrixeinträge s_{ij} der vom Webgraphen eindeutig bestimmten stochastischen Matrix $S \in M_n(\mathbb{C})$:

$$\begin{aligned} s_{ij} &= \tfrac{1}{|O_j|} \quad falls\ j \in I_i \ , \\ s_{ij} &= 0 \quad falls\ j \notin I_i\ und\ O_j \neq \emptyset \ , \\ s_{ij} &= \tfrac{1}{n} \quad falls\ O_j = \emptyset \ . \end{aligned} \tag{3.1}$$

Die Spaltenstochastizität wird also wieder hergestellt. Diese Modifizierung macht allerdings auch noch aus einem anderen Gesichtspunkt Sinn: Man stelle sich erneut den zufälligen Surfer aus Kapitel 1 vor. Die Einträge s_{ij} geben die Wahrscheinlichkeit an, mit der ein zufälliger Surfer, der auf der Webseite W_j gelandet ist, als nächstes zur Webseite W_i gelangt. Landet der zufällige Surfer auf einer Seite, die keine Outlinks hat, so kommt er allein durch Nutzung der Maus nicht mehr von dieser Seite weg. Er muss also eine neue URL in die Adresszeile des Browsers eintippen. Dabei bieten sich nun eben genau so viele Möglichkeiten, wie es Seiten im Web gibt, also n Möglichkeiten. Nach der Modifizierung ist die nächste Station auf der zufälligen Webreise also mit jeweils gleicher Wahrscheinlichkeit $(\frac{1}{n})$ jede beliebige Seite im Web.

3.2.2 Behebung des Problems des Nicht-Zusammenhangs

Im nächsten Schritt wird das Problem der nicht gegebenen Irreduzibilität behoben. Dies wird dadurch erreicht, dass die positive, spaltenstochastische Matrix S kurzerhand in eine strikt positive und immer noch spaltenstochastische Matrix G umgewandelt wird. Damit sind zwei Fliegen mit einer Klappe geschlagen: Weil die Matrix G strikt positiv ist, ist sie nach Satz 2.2.9 nicht nur irreduzibel, sondern sogar primitiv. Die Google-Matrix G erfüllt also nun alle aus Kapitel 2 erhaltenen PageRank-Voraussetzungen.

In der Praxis sieht dies dann so aus, dass die stochastische Matrix S mit einer stochastischen, strikt positiven Matrix T, bei der etwa $\frac{1}{n}$ in jedem Eintrag steht, konvex kombiniert wird (Konvex-Kombinationen zweier stochastischer Matrizen bleiben stochastisch). Für den Beispiel-Webgraph aus Abbildung 3.3 sieht das dann folgendermaßen aus:

$$\alpha \cdot \underbrace{\begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{7} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{7} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 1 & \frac{1}{2} \\ 0 & 0 & \frac{1}{7} & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & \frac{1}{7} & 0 & 0 & 0 & 0 \end{pmatrix}}_{=:\, Stochastische\ Matrix\ S_{[3.3]}} + (1-\alpha) \cdot \underbrace{\begin{pmatrix} \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \end{pmatrix}}_{=:\, Zufalls-Transformationsmatrix\ T_{[3.3]}} =: G_{[3.3]}$$

mit $0 \leq \alpha < 1$.

Für die Interpretation dieser Konvex-Kombination dient wiederum die Vorstellung des zufälligen Surfers. Es wird hier nun davon ausgegangen, dass sich der zufällige Surfer nicht ausschließlich von den Links durchs Web leiten lässt, sondern auch ab und zu – und zwar nicht nur, wenn er auf einer Webseite ohne Outlinks angekommen ist – eine zufällige URL in die Adresszeile des Browser eingibt. Dabei wäre jede Webseite gleichwahrscheinliches nächstes Ziel der zufälligen Webreise, weshalb die Zufalls-Transformationsmatrix T in jedem Eintrag $\frac{1}{n}$ stehen hat. Der Parameter α in der Konvex-Kombination entscheidet nun über die Gewichtung der Matrizen S und T. Je größer α, desto mehr wird die Matrix S und damit der Einfluss der Links auf die Webseiten-Aufrufe eines Surfers gewichtet. Je kleiner α, desto mehr Gewicht erhält die Zufalls-Transformationsmatrix T, d.h. desto

mehr wird davon ausgegangen, dass sich Surfer rein zufällig durchs Netz bewegen.

Allgemein gilt für die Matrixeinträge g_{ij} der vom Webgraphen eindeutig bestimmten Google Matrix $G \in M_n(\mathbb{C})$ mit $0 \leq \alpha < 1$:

$$\begin{aligned} g_{ij} &= \alpha \cdot \tfrac{1}{|O_j|} + (1-\alpha)\tfrac{1}{n} && falls\ j \in I_i\ , \\ g_{ij} &= (1-\alpha)\tfrac{1}{n} && falls\ j \notin I_i\ und\ O_j \neq \emptyset\ , \\ g_{ij} &= \tfrac{1}{n} && falls\ O_j = \emptyset\ . \end{aligned} \tag{3.2}$$

Google verwendet hier angeblich $\alpha = 0,85$ (siehe auch [LaMey], [BrLei], [Gall], [Wills], [Aust]). Das heißt nichts anderes, als dass der zufällige Surfer, der auf einer Webseite angekommen ist, sich zu 85% von den Links (also den Empfehlungen des jeweiligen Webautors) den Weg durchs WWW weisen lässt, und ansonsten nach dem Zufallsprinzip ein neue Webadresse in die Adresszeile des Browsers eintippt. Warum sich diese Parameter-Wahl für Google als optimal herausgestellt hat, soll im nächsten Kapitel untersucht werden. Zunächst soll hier aber noch das gewählte Beispiel zu Ende gerechnet werden.

Mit $\alpha = 0,85$ gilt für die Google Matrix $G_{[3.3]}$ des Webgraphen aus Abbildung 3.3:

$$G_{[3.3]} = \alpha \cdot S_{[3.3]} + (1-\alpha) \cdot T_{[3.3]} = \begin{pmatrix} \frac{3}{140} & \frac{32}{105} & \frac{1}{7} & \frac{25}{56} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{32}{105} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{32}{105} & \frac{32}{105} & \frac{1}{7} & \frac{25}{56} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{32}{105} & \frac{32}{105} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{61}{70} & \frac{25}{56} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{61}{70} & \frac{3}{140} & \frac{25}{56} \\ \frac{3}{140} & \frac{3}{140} & \frac{1}{7} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} & \frac{3}{140} \end{pmatrix}$$

3.3 Berechnung des PageRank-Vektors

Für die Google-Matrix G existiert also nach Theorem 2.1.13 ein eindeutiger (normierter), strikt positiver Fixvektor $x^\star$ – der PageRank-Vektor! Nach Theorem 2.2.7 gilt zudem:

$$\forall x_0 \in \mathbb{C}^n\ mit\ x_0 \geq 0\ und\ \|x_0\|_1 = 1\ :\quad \lim_{k\to\infty} G^k x_0 = x^\star \quad mit \quad x^\star = Gx^\star\ .$$

Für den Beispiel-Webgraphen aus Abbildung 3.3 mit zugehöriger Google-Matrix $G_{[3.3]}$ heißt das für einen beliebigen Vektor $x_0 \in \mathbb{C}^7$ mit $x_0 \geq 0$ und $\|x_0\|_1 = 1$:

$$\lim_{k\to\infty} (G_{[3.3]})^k x_0 = x^\star_{[3.3]} \quad mit \quad x^\star_{[3.3]} = G_{[3.3]} x^\star_{[3.3]}\ .$$

Sei hier etwa $x_0 = (1,0,0,0,0,0,0)^T$, dann setze $x_k = G_{[3.3]}x_{k-1}$ und lasse $k \to \infty$. Dann ist

$$x_1 = G_{[3.3]}x_0 = \begin{pmatrix} 0.021429 \\ 0.304762 \\ 0.304762 \\ 0.304762 \\ 0.021429 \\ 0.021429 \\ 0.021429 \end{pmatrix}, \qquad x_2 = G_{[3.3]}x_1 = \begin{pmatrix} 0.274308 \\ 0.064507 \\ 0.280380 \\ 0.150856 \\ 0.085757 \\ 0.085757 \\ 0.058435 \end{pmatrix}, \qquad \dots \textit{ usw. } \dots,$$

$$x_{41} = G_{[3.3]}x_{40} = \begin{pmatrix} 0.081606 \\ 0.057267 \\ 0.104727 \\ 0.073493 \\ 0.324381 \\ 0.324381 \\ 0.034145 \end{pmatrix} = x_{42} = x_{43} = \ldots = \lim_{k\to\infty} (G_{[3.3]})^k x_0 = x^\star_{[3.3]} .$$

Für die Google-Matrix $G_{[3.3]}$ liefert die Potenzen-Methode bzw. das Iterationsverfahren also bei einer numerischen Berechnungsgenauigkeit von $\frac{1}{2} \cdot 10^{-6}$ nach 41 Iterationen ein stabiles Resultat – den PageRank-Vektor $x^\star_{[3.3]}$.

Die Koordinaten von $x^\star_{[3.3]}$ enthalten die einzelnen PageRank-Werte der Webseiten W_1 bis W_7 aus Abbildung 3.3.

Demnach erhalten die Seiten W_5 und W_6 die höchsten Suchanfragen-unabhängigen Wichtigkeitswerte. Dies lässt sich mit geschärftem Blick auf Abbildung 3.3 dadurch erklären, dass sich die beiden Seiten jeweils mit ihrem einzigen Outlink gegenseitig exklusiv ihren gesamten Wichtigkeitswert zuschanzen und dass sich beide zudem die Stimme eines unabhängigen weiteren Webautors (W_7) teilen.

Trotz der Nachvollziehbarkeit des PageRank-Berechnungsprinzips ist es schwer vorstellbar, wie Googles Rechenzentren die Potenzierung der Google-Matrix mit ihren mehreren Milliarden Zeilen und Spalten meistern können. Die Tatsache, dass der wichtigste Parameter des PageRank-Verfahrens – die Linkmatrix A – eine sogenannte „sparse matrix“ [LaMey] (d.h. die Einträge der Matrix sind größtenteils Nullen, weil Webautoren im Durchschnitt weniger als 10 Links setzen) ist, ist für den Speicher- und Rechenaufwand des Verfahrens von Vorteil. Dennoch ist die Berechnung des PageRank-Vektors eine durchaus beeindruckende Leistung, die angeblich mehrere Tage in Anspruch nimmt.

3.4 Die Parameter der Google-Matrix

Der PageRank-Vektor $x^\star$ ist der strikt positive, auf 1 normierte eindeutige Fixvektor der wie in (3.2) eindeutig bestimmten Google Matrix

$$G = \alpha(A + D) + (1 - \alpha)T,$$

wobei $\alpha = 0,85$, $G \in M_n(\mathbb{C})$ und – mit der Notation und den Definitionen aus Kapitel 1 –

a_{ij} *mit* $\left\{ \begin{array}{lll} a_{ij} = \frac{1}{|O_j|} & falls & j \in I_i \\ a_{ij} = 0 & sonst. & \end{array} \right\}$ die Einträge der Linkmatrix A,

d_{ij} *mit* $\left\{ \begin{array}{lll} d_{ij} = \frac{1}{n} & falls & O_j = \emptyset \\ d_{ij} = 0 & sonst. & \end{array} \right\}$ die Einträge der Matrix D und

t_{ij} *mit* $t_{ij} = \frac{1}{n} \quad \forall\, i, j.$ die Einträge der Matrix T sind.

Da das Aussehen der Matrix D abhängig ist vom Aussehen der Linkmatrix A (hat die Seite W_j keine Outlinks, so stehen in der j-ten Spalte von A nur Nullen und dementsprechend sind alle Einträge der j-ten Spalte von D $= \frac{1}{n}$), sollen im Folgenden mögliche Modifikationen der drei entscheidenden Parameter A, α und T und deren Auswirkungen auf das PageRank-Verfahren untersucht werden.

3.4.1 Die Linkmatrix A

Der bisher diskutierte Ansatz sieht ein einheitliches Gewichtungsprinzip der Nicht-Nulleinträge von A ($\frac{1}{|O_j|}$) vor. Das heißt, dass der in Kapitel 1 vorgestellte zufällige Surfer mit jeweils gleicher Wahrscheinlichkeit einem der existierenden Outlinks einer Seite folgt, wenn er nur die Maus zum Surfen benutzen darf. Beäugt man diesen Ansatz kritisch, so könnte man anzweifeln, ob er tatsächlich ein realistisches Ranking produziert. Wäre es nicht sinnvoller, den „zufälligen Surfer“ durch einen „intelligenten Surfer“ ([LaMey], S.48) bzw. einen „menschlicheren“ Surfer zu ersetzen, der sich beim Verlassen einer Webseite beispielsweise von der Reihenfolge der angebotenen Outlinks oder der Linkbeschreibung – also Angaben über den Inhalt der verlinkten Seiten – beeinflussen lässt?

Man könnte etwa, nur die Reihenfolge in Betracht ziehend, der Auffassung sein, dass der intelligente Surfer dem erstgenannten Outlink einer Seite doppelt so wahrscheinlich folgt wie den anderen Outlinks. Gehen wir nun im Beispiel des Webgraphen aus Abbildung 1.3 davon aus, dass der Webautor von Seite W_1 den Outlink zu Seite W_4 an erster Stelle nennt und die Webautoren der Seiten W_2 und W_4 jeweils den Outlink zu Seite W_3 zuerst anführen (W_3 hat nur einen Outlink), so würde sich die

$$\text{ursprüngliche Linkmatrix } \begin{pmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{pmatrix} \text{ ändern in } \begin{pmatrix} 0 & 0 & 1 & \frac{1}{3} \\ \frac{1}{4} & 0 & 0 & 0 \\ \frac{1}{4} & \frac{2}{3} & 0 & \frac{2}{3} \\ \frac{1}{2} & \frac{1}{3} & 0 & 0 \end{pmatrix}.$$

Nach gleichem Prinzip könnten auch z.B. die Länge des Anchortexts oder die inhaltliche Ähnlichkeit der verlinkten Seiten mit in die Gewichtung der Spalteneinträge eingehen. Wichtigste Maßgabe bleibt dabei natürlich, dass die Spaltensumme in jedem Falle 1 ergeben muss. Übereinstimmenden Quellen zu Folge (z.B. [LaMey], [BrLei]) gingen Larry Page und Sergey Brin ursprünglich vom Ansatz mit den gleichgewichteten Outlinks, also dem Prinzip des zufälligen Surfers, aus. Inwieweit Google den Surfer heutzutage als „intelligenter“ oder „menschlicher“ interpretiert, d.h. ob, und wenn ja, welche Modifikationen an der ursprünglichen Linkmatrix vorgenommen wurden, bleibt Googles Geheimnis.

3.4.2 Der Einfluss von α

Aufgrund des Erfolgs des Google-Algorithmus kann man durchaus sagen, dass sich die Wahl des Parameters α $(= 0,85)$ absolut bewährt hat. Grund genug, dessen Einfluss auf das PageRank-Verfahren etwas genauer zu untersuchen. Am Beispiel des Webgraphen aus Abbildung 3.3 mit der zugehörigen Google-Matrix $G_{[3.3]}$ lässt sich einfach feststellen, welchen Einfluss α auf das Verfahren und die Rankingwerte nimmt. Man betrachte die folgende Tabelle, in der die Ergebnisse der PageRank-Vektor-Berechnung für verschiedene α-Werte stehen:

$\alpha =$	0,85	0,95	0,5	0,1
Iterationen bis Stabilität:	41	60	17	7
$x^{\star}_{[3.3]} =$	$\begin{pmatrix} 0.081606 \\ 0.057267 \\ 0.104727 \\ 0.073493 \\ 0.324381 \\ 0.324381 \\ 0.034145 \end{pmatrix}$	$\begin{pmatrix} 0.039116 \\ 0.026519 \\ 0.051503 \\ 0.034917 \\ 0.416906 \\ 0.416906 \\ 0.014133 \end{pmatrix}$	$\begin{pmatrix} 0.129870 \\ 0.103896 \\ 0.151515 \\ 0.121212 \\ 0.205628 \\ 0.205628 \\ 0.082251 \end{pmatrix}$	$\begin{pmatrix} 0.14218 \\ 0.13541 \\ 0.14692 \\ 0.13992 \\ 0.15245 \\ 0.15245 \\ 0.13067 \end{pmatrix}$
Numerische Berechnungsgenauigkeit: $\frac{1}{2} \cdot 10^{-6}$				

Die in der Tabelle festgehaltenen Ergebnisse legen folgende Interpretation nahe:
Je kleiner α,

- desto höher die Konvergenzgeschwindigkeit, d.h. desto weniger Iterationen sind bis zur Konvergenz des Verfahrens nötig,
- aber desto undeutlicher werden auch die Rankingwert-Unterschiede (da die Linkmatrix bei kleinerem α – und damit größerem Anteil der Zufallsmatrix T in der Konvex-Kombination – stärker verfälscht wird).

Natürlich soll der Rechenaufwand bei der Berechnung des PageRank-Vektors möglichst minimiert werden. Eine hohe Konvergenzgeschwindigkeit ist also wünschenswert. Bedenkt man allerdings die Größe des WWW und die Anzahl der indizierten Webseiten, die alle mit einem PageRank-Wert versehen werden müssen, so wird klar, dass auch sehr feine Rankingwert-Unterschiede von Bedeutung sein können. Verfälscht man hier die Linkmatrix zu stark (durch Wahl eines sehr kleinen α), so könnte eventuell auch das Ranking zu stark verfälscht werden. Es gilt also, eine optimale Zwischenlösung zu finden.

Stabilität des Rankings bei verschiedenen α-Werten

Obwohl sich die Rankingwerte bei kleinem α immer mehr annähern, bleibt das Ranking in der obigen Tabelle bemerkenswerterweise bei allen vier α-Werten gleich, was darauf zurückzuführen ist, dass der zu Grunde liegende Webgraph aus Abbildung 3.3 nur 7 Webseiten enthält. Das Ranking bleibt hier – bei der angegebenen Berechnungsgenauigkeit – noch stabil. Spannender wird es, wenn man den Einfluss von α auf das Ranking größerer Webs untersucht (siehe Abb. 3.4):

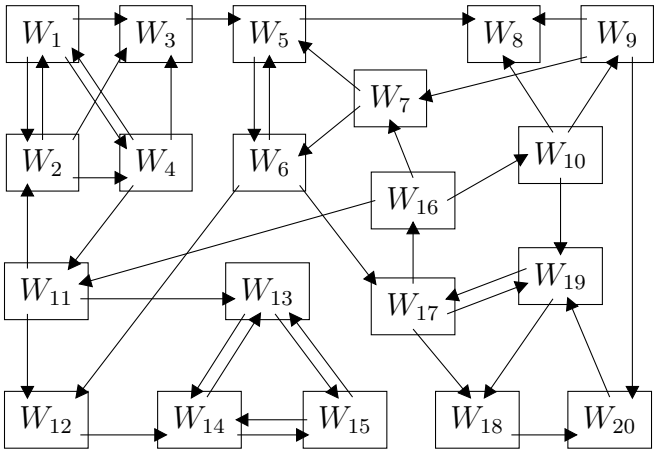


Abbildung 3.4: Web aus 20 Seiten ↑ , ↓ zugehörige PageRank-Werte für verschiedene α

	$\alpha = 0,95$		$\alpha = 0,85$		$\alpha = 0,5$		$\alpha = 0,1$	
Seite	PR-Wert	Rang	PR-Wert	Rang	PR-Wert	Rang	PR-Wert	Rang
W_1	0.0102138	17	0.021396	16	0.039495	13	0.048488	13
W_2	0.0107904	16	0.021630	15	0.039392	15	0.048486	15
W_3	0.0134482	13	0.027458	11	0.046077	11	0.050104	9
W_4	0.0102138	17	0.021396	16	0.039495	13	0.048488	13
W_5	0.0289840	8	0.052565	8	0.067522	4	0.054365	2
W_6	0.0227664	9	0.040182	9	0.052598	8	0.050393	8
W_7	0.0114451	15	0.020321	18	0.037481	17	0.048380	17
W_8	0.0223676	10	0.040115	10	0.053879	7	0.051095	6
W_9	0.0065318	20	0.013851	20	0.031713	20	0.046816	20
W_{10}	0.0093767	19	0.016397	19	0.032195	19	0.046819	19
W_{11}	0.0126111	14	0.022459	14	0.038778	16	0.048435	16
W_{12}	0.0147654	12	0.026953	12	0.041576	12	0.048550	12
W_{13}	0.1900128	2	0.124034	2	0.068765	3	0.052164	4
W_{14}	0.1968152	1	0.135646	1	0.080225	1	0.055250	1
W_{15}	0.1873053	3	0.119568	3	0.063594	5	0.050626	7
W_{16}	0.0183608	11	0.025384	13	0.035090	18	0.046909	18
W_{17}	0.0467315	7	0.057101	7	0.052458	9	0.049618	10
W_{18}	0.0543204	6	0.061895	6	0.052435	10	0.049592	11
W_{19}	0.0757041	4	0.085909	4	0.069381	2	0.053648	3
W_{20}	0.0572355	5	0.065740	5	0.057850	6	0.051775	5
Numerische Berechnungsgenauigkeit: $\frac{1}{2} \cdot 10^{-6}$								

Insgesamt kann man sagen, dass das Ranking auch bei dem gewählten 20×20-Beispiel erstaunlich stabil bleibt. Allerdings lässt sich aus der Tabelle in Abbildung 3.4 ablesen, dass sich zumindest die Ranking-Positionen mancher Webseiten für verschiedene α-Werte verändern. Vergleicht man das Ranking der Seiten für $\alpha = 0,95$ mit dem Ranking für $\alpha = 0,85$, so stellt man fest, dass sich die Position im Ranking nur für 6 Seiten minimal ändert, wobei die größte Änderung im Abstieg der Seite W_7 von Rang 15 auf Rang 18 besteht. Das Ranking für $\alpha = 0,1$ ändert sich dagegen im Vergleich zu $\alpha = 0,85$ auf 14 Positionen, teilweise sehr deutlich. Seite W_5 verbessert sich z.B. um ganze 6 Ränge von Platz 8 auf Platz 2.

Der Zusammenhang zwischen α und der Konvergenzgeschwindigkeit

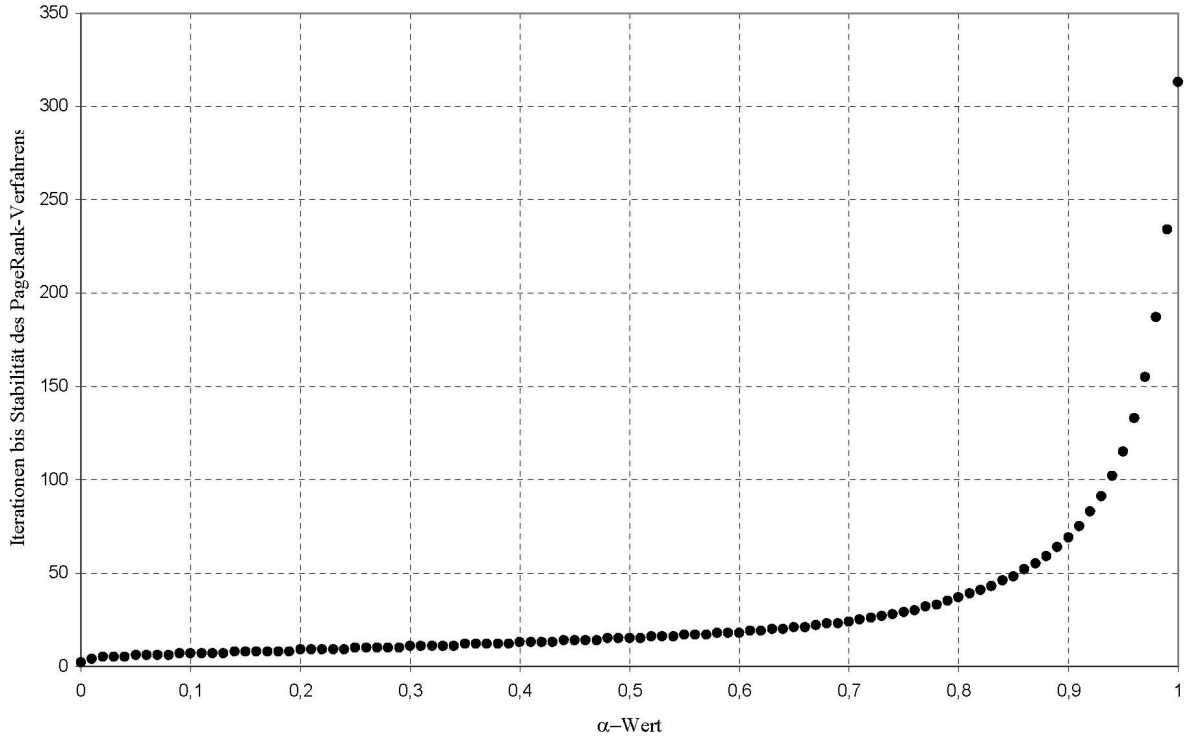


Abbildung 3.5: α-abhängige Konvergenzgeschwindigkeiten des PageRank-Verfahrens für das Beispiel-Web aus Abbildung 3.4 bei einer Berechnungsgenauigkeit von $\frac{1}{2} \cdot 10^{-6}$

Die Ergebnisse aus dem Diagramm in Abbildung 3.5 bestärken die Vermutung, dass die Konvergenzgeschwindigkeit des PageRank-Verfahrens von α abhängt. Je kleiner α ist, desto weniger Iterationen werden demnach auch in diesem Beispiel benötigt, um ein stabiles Resultat zu erhalten. Taher H. Haveliwala und Sepandar D. Kamvar haben sich in ihrer Arbeit [HaKa] intensiver mit diesem Zusammenhang beschäftigt und bewiesen folgendes Theorem, das seinerseits den Beweis zur geäußerten Vermutung liefert:

$$\alpha =: \lambda_2 \textit{ ist ein Eigenwert der Google-Matrix } G,$$
$$\textit{wobei } |\lambda_i| \leq |\lambda_2| \textit{ für alle } \lambda_i \textit{ außer } \lambda_1 = 1 = r(G) \textit{ gilt.}$$

Damit ist klar, dass die Folgen $z_{\lambda_i,\nu}(k) := \binom{k}{\nu}\lambda_i^{k-\nu} \ \forall \ \lambda_i \neq 1$ für $k \to \infty$ umso schneller gegen 0 – und damit J^k umso schneller gegen P_1 – konvergieren, je kleiner α ist.

$\alpha = 0,85$ – **Die optimale Wahl?**

Die Punktkurve in Abbildung 3.5 hat bis zum x-Achsenbereich um $\alpha = 0,85$ eine relativ flache Steigung, während die Zahl der benötigten Iterationen bis zum Erhalt eines stabilen Resultats für größere α-Werte sehr schnell und sehr stark ansteigt. Diese Beobachtung legt die Vermutung nahe, dass die Konvergenzgeschwindigkeit des PageRank-Verfahrens für $\alpha = 0,85$ für die Google-Entwickler gerade noch akzeptabel ist, während der Rechenaufwand für größere α-Werte nicht mehr – oder nicht im gewünschten Zeitrahmen – zu bewältigen ist.

Da man α und damit den Einfluss der Linkstruktur als Indikator für das Ranking natürlich so groß wie möglich haben will, kann die Wahl $\alpha = 0,85$ also als optimaler Kompromiss zwischen „Effizienz und Effektivität“ [LaMey] bezeichnet werden.

3.4.3 Die Zufalls-Transformationsmatrix T

Die Einträge der Linkmatrix A geben die Wahrscheinlichkeiten an, mit welcher ein zufälliger oder intelligenter Surfer (siehe 3.4.1) das WWW über die angebotenen Links (d.h. – bildlich gesprochen – die Maus benutzend) bereist. Die Matrix T deckt dagegen den Fall ab, dass der Surfer die Links auf den besuchten Webseiten komplett ignoriert und sich nur mit der Tastatur – also immer wieder neue Webadressen in die Adresszeile des Browsers eingebend – durchs WWW bewegt. Der bisher diskutierte Ansatz sieht hier zunächst ebenfalls das Zufallsprinzip vor. Die Einträge t_{ij} von T mit $t_{ij} = \frac{1}{n} \ \forall \ i,j$ besagen, dass der zufällige Surfer alle existierenden Webadressen mit jeweils gleicher Wahrscheinlichkeit in die Adresszeile eintippt.

Versieht man den imaginären zufälligen Surfer wie zuvor mit etwas menschlicheren Zügen, so liegt die Einsicht nahe, dass eine menschlicher Surfer nicht allwissend ist, und daher nicht einmal annähernd sämtliche Adressen des WWW kennt. Außerdem hat ein menschlicher Surfer bestimmte Interessen und Vorlieben. Ein typischer Fußballfan beispielsweise wird weit weniger wahrscheinlich eine Seite mit politischem Inhalt ansteuern, als eine der zahlreichen ihm bekannten Sportseiten. Sei also im Web aus Abbildung 1.3 beispielsweise $W_1 = www.kicker.de$, $W_2 = www.ard.de$, W_3 eine Seite über die Kommunalpolitik eines norddeutschen Provinzortes und $W_4 = www.sport1.de$. Dann könnte die auf einen VfB Stuttgart-Fan zugeschnittene Transformations Matrix z.B. folgendermaßen aussehen:

$$\begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & 0 & 0 \\ \frac{3}{10} & \frac{3}{10} & \frac{3}{10} & \frac{3}{10} \end{pmatrix}$$

Die notwendige mathematische Eigenschaft – die Spaltenstochastizität – bliebe dabei erhalten und die Wahrscheinlichkeits-Einträge der Matrix und damit letztendlich auch der PageRank-Vektor würden in gewissem Maße den Interessen des VfB-Fans entsprechen: Dass alle Einträge in der dritten Zeile = 0 sind, könnte bedeuten, dass er die Kommunalpolitik-Seite W_3 gar nicht kennt und deren Adresse dementsprechend auch nie in die Browserzeile eintippen kann. Am wahrscheinlichsten ist es dagegen, dass er *www.kicker.de* ansteuert ($\frac{1}{2}$), gefolgt von *www.sport1.de* ($\frac{3}{10}$) und *www.ard.de* ($\frac{1}{5}$).

Eine derartige die Interessen des Users berücksichtigende Suchoption brachte Google im März 2004 als Beta-Version auf den Markt, wobei noch ausdrücklich darauf hingewiesen wird, dass sich die in den Google Labs angebotenen Beta-Dienste noch in der Entwicklung befinden und daher unter Umständen zeitweise nicht funktionieren. Der entsprechende Link unter http://labs.google.de/ wird mit *Personalisierte Suche* bezeichnet. Natürlich kann Google nicht für jeden User individuell die riesige Google-Matrix und einen persönlichen PageRank-Vektor berechnen. Statt dessen wird die Möglichkeit geboten, sich ein Profil zu erstellen, in dem man einige von Google angebotene Interessens-Kategorien verschieden gewichtet. Die „personalisierte“ Suche verwendet dann einen dieser Gewichtung entsprechenden, angepassten PageRank-Vektor.

Anstatt die Transformationsmatrix T zu personalisieren, kann man die Übergangsswahrscheinlichkeiten von einer Seite zur anderen auch von der inhaltlichen Nähe der beiden Seiten abhängig machen. Ohne von vorne herein die Interessen des zufälligen oder intelligenten Modell-Surfers festzulegen, würde doch die Einschätzung Sinn machen, dass ein Surfer, der sich im zuvor beschriebenen Beispiel-Web auf der Kicker-Seite befindet, sehr viel wahrscheinlicher als nächstes zu einer inhaltsverwandten Seite wie *www.sport1.de* wechselt als auf die Seite über norddeutsche Kommunalpolitik. Jemand, der sich für Politik interessiert und sich gerade auf der Kommunalpolitik-Seite befindet, wird dagegen als nächstes vielleicht eher die ARD-Seite aufrufen als die Kicker-Seite. Allerdings ist diese Inhalts-Beziehung und damit die Matrix nicht zwangsläufig symmetrisch. Auch wenn ein Besucher der Politik-Seite in unserem Beispiel als nächstes mit höchster Wahrscheinlichkeit *www.ard.de* aufruft, heißt das nicht, dass ARD-Seitenbesucher mit gleicher Wahrscheinlichkeit zur Politikseite wechseln.

Geht man zudem von einem Mindestmaß an Intelligenz des Surfers aus, so sollte man berücksichtigen, dass die Wahrscheinlichkeit, dass ein sich z.B. auf *www.ard.de* befindender Surfer die gleiche Adresse nochmals eingibt, fast Null ist (sie darf nicht gleich Null sein, weil eine strikt positive Google-Matrix produziert werden soll). Eine nach diesem Prinzip inhaltsorientierte Transformationsmatrix könnte in unserem Beispiels etwa folgendermaßen gestaltet sein:

$$\begin{pmatrix} \frac{1}{100} & \frac{2}{5} & \frac{9}{100} & \frac{1}{2} \\ \frac{3}{10} & \frac{1}{100} & \frac{7}{10} & \frac{2}{5} \\ \frac{9}{100} & \frac{7}{50} & \frac{1}{100} & \frac{9}{100} \\ \frac{3}{5} & \frac{9}{20} & \frac{1}{5} & \frac{1}{100} \end{pmatrix}$$

Die Frage, inwieweit Googles Zufalls-Transformationsmatrix T tatsächlich noch strikt nach dem ursprünglichen Zufallsprinzip oder doch eher die inhaltliche Nähe der verschiedenen Seite berücksichtigend gestaltet ist, können wiederum nur die Google-Ingenieure beantworten. Im eigenen Interesse werden sie uns diese Antwort allerdings so lange wie möglich schuldig bleiben.

Da es Gerüchten zu Folge mehr als 100 Parameter im Google-Algorithmus geben soll und Google weiterhin um Innovation und Verbesserung bemüht ist, ist es allerdings unrealistisch, dass Google nach wie vor vom Prinzip des strikt zufälligen Surfers ausgeht. Schließlich war und bleibt es Googles erklärtes Ziel, ein realistisches Ranking zu produzieren, welches weitestgehend menschlichen Rankings entspricht. Es ist also davon auszugehen, dass den Google-Entwicklern die obigen Gedanken nicht gänzlich unbekannt sind und daher auch die Transformationsmatrix T im Laufe der Weiterentwicklungen und Verbesserungen nicht von Modifikationen verschont blieb.

4 PageRank ohne Spektraltheorie?

Thema dieser Arbeit ist die Mathematik hinter PageRank. In Kapitel 2 wurde gezeigt, wie spektraltheoretische Argumente die entscheidende Aussage für das PageRank-Verfahren liefern:

$$\forall x_0 \in \mathbb{C}^n \; mit \; x_0 \geq 0 \; und \; \|x_0\|_1 = 1 \; : \quad \lim_{k\to\infty} G^k x_0 = x^\star \quad mit \quad x^\star = Gx^\star \tag{4.1}$$

(wobei G die Google-Matrix ist, die u.a. insbesondere positiv und spaltenstochastisch ist).

Die gleiche Aussage erhält man jedoch auch ganz ohne Spektraltheorie und erstaunlich schnell und direkt aus dem Banachschen[4] Fixpunktsatz. Allerdings lässt sich dieses alternative Argument, wie wir gleich sehen werden, nicht mit gewissen Parameter-Variationen vereinbaren.

4.1 Banachscher Fixpunktsatz

Sei (M, d) ein vollständiger metrischer Raum, $\varphi : M \to M$ eine kontrahierende Abbildung, d.h.

$$\exists\, 0 \leq \alpha < 1, \text{ so dass } \forall\, x, y \in M : \quad d(\varphi(x), \varphi(y)) \leq \alpha \cdot d(x, y).$$

Dann existiert genau ein Fixpunkt von φ, d.h.

$$\exists!\, x^\star \in M \quad \text{mit} \quad \varphi(x^\star) = x^\star.$$

Für alle $x_0 \in M$ konvergiert die rekursiv definierte Folge $x_k := \varphi(x_{k-1})$ gegen $x^\star$.

Für den Beweis des Satzes und die anschließende Anwendung siehe [WHK], S. 421-426.

Anwendung auf PageRank

Man betrachte die Menge

$$M = \left\{ (x_1, \ldots, x_n) \;\mid\; x_i \geq 0, \; \sum_{i=1}^{n} = 1 \right\}$$

mit der von der Norm induzierten Metrik $d(x, y) = \|x - y\|_1$. Dann ist M offensichtlich abgeschlossen und damit als Teilraum von $\mathbb{R}^n$ vollständig.

[4] Stefan Banach (1892-1945)

Wir müssen nun nur noch zeigen, dass die wie in (3.2) eindeutig bestimmte Google-Matrix G eine kontrahierende Abbildung auf M definiert und erhalten dann mit dem Banachschen Fixpunktsatz den eindeutig bestimmten positiven Fixvektor von G mit Norm 1.

Wegen Linearität gilt

$$Gx - Gy = \alpha \cdot S(x - y) + (1 - \alpha)(Tx - Ty) \quad \forall\, x, y \in M,$$

wobei $0 \leq \alpha < 1$,

S die wie in (3.1) eindeutig durch den Webgraphen bestimmte stochastische Matrix

und $T = \frac{1}{n} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$ die Zufalls-Transformationsmatrix ist.

Es ist dann $Tx = Ty = (\frac{1}{n}, \ldots, \frac{1}{n})\ \forall\, x, y \in M$ und damit

$$Gx - Gy = \alpha \cdot S(x - y) \quad \forall\, x, y \in M.$$

Folglich ist die von G definierte Abbildung wegen

$$\|Gx - Gy\|_1 = \|\alpha S(x - y)\| \leq \alpha \underbrace{\|S\|_1}_{=1} \|x - y\|_1 = \alpha \|x - y\|_1$$

eine kontrahierende Abbildung im Sinne des Banachschen Fixpunktsatzes.

4.2 Perron-Frobenius vs. Banach

Der vorangegangene Abschnitt beschreibt auf etwas mehr als einer Seite scheinbar die komplette für das PageRank-Verfahren benötigte Mathematik. Der Banachsche Fixpunktsatz liefert also ein gleichstarkes Argument wie die von Seite 16 bis 29 (Kapitel 2) aufwendig bemühte Spektraltheorie – so könnte man meinen. Wozu also der ganze Aufwand mit der Perron-Frobenius Argumentation, wenn es uns Banach so einfach macht?

Die Antwort ist nicht offensichtlich. Nur bei genauer Untersuchung kann man den Vorteil der spektraltheoretischen Argumentation erkennen: Sie ist flexibler! Modifiziert man beispielsweise die Transformationsmatrix T wie in Abschnitt 3.4.3 beschrieben so, dass die Einträge t_{ij} und damit die Übergangswahrscheinlichkeit von Seite W_j zu Seite W_i von der inhaltlichen Nähe der beiden Seiten abhängen, so ist das Banachsche Argument nicht mehr anwendbar: Da die Spalten von T nicht mehr gleich aussähen, würde nicht $Tx = Ty\ \forall\, x, y \in M$ (siehe Abschnitt 4.1, Anwendung auf PageRank) gelten.

Das Ergebnis eines imaginären Showdowns zwischen Perron-Frobenius und Banach hieße also – ohne den Wert der zahlreichen Arbeiten Banachs auf verschiedenen Gebieten der Mathematik geringer schätzen zu wollen – in diesem speziellen Anwendungsgebiet oder, bildlich gesprochen, auf diesem speziellen Spielfeld:

1 zu 0 nach Verlängerung für die Spielgemeinschaft Perron-Frobenius.

Literaturverzeichnis

[Aust] D. Austin, How Google Finds Your Needle in the Web's Haystack. *Feature Column. Monthly Essays on Mathematical Topics*. American Mathematical Society, 2006. http://www.ams.org/featurecolumn/archive/pagerank.html#2. Zugriff am 6.4.2007

[BNS] A. Bátkai, R. Nagel, U. Schlotterbeck, *An Invitation to Positive Matrices*. Budapest/Tübingen, 2006.

[BrLei] K. Bryan, T. Leise, The $25,000,000,000 Eigenvector: The Linear Algebra behind Google. *SIAM Review* 48 (3), S. 569-81. 2006.

[Fisch] G. Fischer, *Lineare Algebra*. Vieweg & Sohn Verlagsgesellschaft mbH, Braunschweig/Wiesbaden, 1997.

[FN07] N.N., „Firma, die man bei Google nicht findet, existiert nicht". *Fränkische Nachrichten* vom 6. März 2007, S. 13.

[Gall] P.F. Gallardo, Google's secret and linear algebra. *Newsletter of the European Mathematical Society* 63, S. 10-15. European Mathematical Society, Zürich, 2007.

[Goog1] *http://www.google.de/intl/de/options/*, Google-Suche: Mehr, mehr, mehr... . Zugriff am 6.4.2007.

[Goog2] *http://www.google.com/corporate/history.html*, Google Coorporate Information: Google Milestones. Zugriff am 6.4.2007.

[Goog3] *http://www.google.com/corporate/tech.html*, Unternehmensbezogene Informationen zu Google: Technologie. Zugriff am 6.4.2007.

[Goog4] *http://www.google.de/why_use.html*, Gründe, Google zu benutzen. Zugriff am 6.4.2007.

[HaKa] T.H. Haveliwala, S.D.Kamvar, *The second eigenvalue of the Google matrix*. Technical Report. Stanford University, 2003.

[HB06] N.N., Der Mann des Jahres: Google-Chef Eric Schmidt im Exklusiv-Interview. *Handelsblatt* (Nr. 248) vom 22. Dezember 2006.

[Kaim] S. Kaim, *Google zeigt mich, also bin ich*. 45minütige Dokumentation. ARTE France, Frankreich, 2006. Ausgestrahlt am 12. Dezember 2006 auf ARTE im Rahmen der Sendung *Generation Ahnungslos*.

[LaMey] A.N. Langville, C.D. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Princeton University Press, Princeton, 2006.

[MacC] C.R. MacCluer, The many proofs and applications of Perron's theorem. *SIAM Review* 42 (3), S. 487-498. 2000.

[Minc] H. Minc, *Nonnegative Matrices*. John Wiley and Sons Interscience, New York, 1988.

[RobFo] D.F. Robinson, L.R. Foulds, *Digraphs: Theory and Techniques*. Gordon and Breach Science Publishers, New York/London/Paris, 1980.

[Schae] H.H. Schaefer, *Banach Lattices and Positive Operators*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellung, Band 215, Springer-Verlag, Berlin/Heidelberg/New York, 1974.

[Schö] V.C. Schöch, *Die Suchmaschine Google*. Seminararbeit. Institut für Informatik, Freie Universität Berlin, 2001.

[WHK] M. Wolff, P. Hauck, W. Küchlin, *Mathematik für Informatik und Bioinformatik*. Springer-Verlag, Berlin, 2004.

[Wills] R.S. Wills, Google's PageRank. The Math Behind the Search Engine. *The Mathematical Intelligencer* 28 (4), S. 6-11. Springer-Verlag, 2006.

[Wint] M. Winter, *Spektraltheorie positiver Matrizen und Asymptotik von Operatorpotenzen*. Zulassungsarbeit der Fakultät für Mathematik und Physik an Eberhard-Karls-Universität Tübingen, 2002.