Archenhold Gymnasium

Tutorium Kneiske

Betreuende Lehrkraft: Frau Kaiser

Schuljahr 2022/2023

Der PageRank-Algorithmus – Wie könnte eine geheime Erweiterung des PageRank-Algorithmus von Google aussehen?

Besondere Lernleistung im Fach Informatik
mit Bezug auf das Unterrichtsfach Mathematik
von

Floyd Wollert

Abgabe am 19.12.2022

Inhaltsverzeichnis

Vorwort

In dieser Arbeit habe ich mich entschieden, die Standardversion des PageRank-Algorithmus anstelle der neueren Version zu verwenden. Diese Entscheidung beruht auf mehreren Faktoren. Erstens habe ich mich dafür entschieden, weil die Standardversion des Algorithmus weiter verbreitet und besser etabliert ist. Das bedeutet, dass es eine größere Menge an verfügbaren Daten und Studien gibt, die sich mit dieser Version des Algorithmus befassen. Zweitens habe ich mich dafür entschieden, weil ich glaube, dass die Standardversion des Algorithmus genau genug ist, um meine Forschung durchzuführen. Auch wenn die neuere Version einige Verbesserungen aufweist, habe ich mich dafür entschieden, dass diese Verbesserungen für meine Arbeit nicht zweckdienlich sind. Drittens glaube ich, dass die Verwendung der Standardversion des Algorithmus den Vergleich der Ergebnisse meiner Arbeit mit den Ergebnissen früherer Studien erleichtert. Durch die Verwendung der Standardversion kann ich die Ergebnisse meiner Arbeit leichter in den Kontext früherer Untersuchungen einordnen.

Motivation

Das Thema PageRank-Algorithmus ist von großer Bedeutung, da er dazu beiträgt, das Suchverhalten und die Nutzung des Internets zu verbessern. Ein tieferes Verständnis dieses Algorithmus kann helfen, die Relevanz von Webseiten für bestimmte Suchanfragen besser zu bestimmen und so die Suche nach relevanten Informationen im Internet zu erleichtern. Eine geheime Erweiterung des PageRank-Algorithmus von Google könnte das Internet in mehrfacher Hinsicht beeinflussen. Einerseits könnte sie dazu beitragen, dass bestimmte Webseiten in den Google-Suchergebnissen besser gefunden werden, was zu mehr Besuchern und damit zu mehr Einnahmen führen könnte. Andererseits könnten andere Websites dadurch benachteiligt werden, was zu weniger Besuchern und weniger Einnahmen führen könnte. Insgesamt könnte eine geheime Erweiterung des PageRank-Algorithmus von Google das Suchverhalten der Nutzer beeinflussen und damit Auswirkungen auf das gesamte Internet haben.

1. Einleitung

Wenn wir heutzutage im Internet nach etwas suchen, wollen wir eine vertrauenswürdige und relevante Webseite möglichst weit oben in den Suchergebnissen finden. Dazu muss eine Suchmaschine in der Lage sein, die Eingaben eines Nutzers zu interpretieren und dann die passenden Webseiten in der richtigen Reihenfolge anzuzeigen. Derzeit gibt es etwa 1,88 Milliarden Webseiten im World Wide Web [4], von denen immer mehrere Tausend auf eine Suchanfrage passen. Daher ist eine solche Reihenfolge im Internet besonders wichtig, denn wir Menschen haben weder die Zeit noch die Lust, Tausende von Seiten zu durchsuchen, bis wir die eine passende Seite finden. Es wurde sogar erwiesen, dass die meisten Webnutzer nicht über die erste Seite der Ergebnisse hinausschauen. In der Tat finden fast 92 % des Google-Datenverkehrs auf Seite 1 statt [60]. Die Effektivität einer Suchmaschine hängt also davon ab, die gewünschten Ergebnisse in einer gut sortierten Reihenfolge auszugeben.

Auch Google beschrieb diese Herausforderung auf ihrer Seite im Jahr 2013:

> *"For every search query performed on Google […] there are thousands, if not millions of web pages with helpful information. Our challenge in search is to return only the most relevant results at the top of the page, sparing people from combing through the less relevant results below. Not every website can come out at the top of the page, or even appear on the first page of our search results. [https://web.archive.org/web/20130731231230/http://www.google.com/competition/howgooglesearchworks.html (2013). Abgerufen am 18. Dezember 2022]*

Ein solches Ranking kann durch den Einsatz von Analysealgorithmen erreicht werden. Eine besondere Art wird im Internet verwendet, nämlich Algorithmen zur Linkanalyse. Der wohl bekannteste dieser Algorithmen ist der PageRank-Algorithmus von Google. Er ist ein wesentlicher Bestandteil der Suchmaschinentechnologie des Unternehmens und wurde 1996 von den Gründern Larry Page und Sergey Brin im Rahmen eines Forschungsprojekts an der Stanford University über eine neue Art von Suchmaschine entwickelt und 1997 patentiert [43]. Der Name des Algorithmus setzt sich aus dem Nachnamen Page, von Larry Page, und der Funktion des Algorithmus „to rank" zusammen.

Der PageRank-Algorithmus trägt dazu bei, die Genauigkeit und Relevanz der Suchergebnisse zu verbessern. Infolgedessen hält Google seinen Marktanteil bei Suchmaschinen weiterhin bei mehr als 90%, während täglich mehr als 3,5 Milliarden Suchanfragen von Google verarbeitet

werden [30]. Dieser große Marktanteil bedeutet, dass Google als Suchmaschine als eines der dominierenden Akteure auf dem Online-Suchmarkt gilt und beschuldigt wurde, seine Macht zu nutzen, um den Wettbewerb zu unterdrücken und den Markt zu kontrollieren [3, 33]. Dies hat einige dazu veranlasst, Google als Quasi-Monopolisten zu betrachten, obwohl das Unternehmen selbst diese Behauptungen bestreitet.

Google könnte also das Internet durch eine geheime Erweiterung seines PageRank-Algorithmus massiv verändern. Um eine solche geheime Erweiterung besser zu verstehen, sollte man zunächst verstehen, wie eine Suchmaschine funktioniert und was genau der PageRank-Algorithmus ist und bewirkt.

2. Web-Suchmaschinen als Information Retrieval-Systeme

Web-Suchmaschinen gehören zum Bereich des Information Retrieval. Information Retrieval ist der Prozess des Auffindens und Abrufens von Informationen aus einer Datensammlung, normalerweise einer Datenbank. Es geht darum, relevante Daten auf der Grundlage der Suchanfrage eines Nutzers zu identifizieren und dann die Ergebnisse so zu organisieren und zu präsentieren, dass sie für den Nutzer nützlich und einfach zu verstehen sind. Internet-Suchmaschinen sind dafür da, das World Wide Web (WWW) zu durchsuchen.

2.1. Bestandteile einer Web-Suchmaschine

Der Suchprozess ist ein vielschichtiger Prozess. Bevor ein Benutzer eine Suchanfrage an eine Suchmaschine senden kann, muss die Suchmaschine zunächst die relevanten Informationen aus dem World Wide Web beschaffen. Das World Wide Web ist kein physischer Ort, sondern ein virtueller Raum, der online existiert. Das Web besteht aus Milliarden miteinander verbundener, mit Hyperlinks versehener Dokumente, Bilder, Videos und anderer digitaler Inhalte, die auf Servern in aller Welt gespeichert sind. Wenn Sie auf eine Webseite zugreifen, sendet Ihr Computer oder Gerät eine Anfrage an den Server, auf dem die Inhalte gespeichert sind und der Server sendet die Inhalte dann an ein Gerät zurück, damit sie in einem Webbrowser angezeigt werden. Auf diese Weise ist das Web ein verteiltes Netz von Informationen, die auf verschiedenen Geräten, die mit dem Internet verbunden sind, abgerufen und angezeigt werden [vgl. 36].

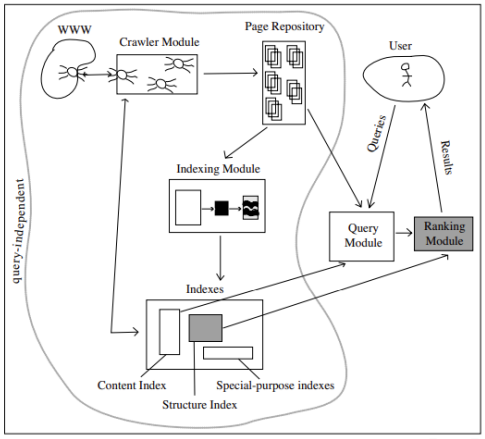


Figure 1.2 Elements of a search engine [36]

Um auf Daten aus dem World Wide Web zugreifen zu können, benötigt jeder Webbrowser sogenannte "Crawler-Module". Dieses Modul enthält Web-Crawler, Internet-Bots, die das World Wide Web durchsuchen. Ein Web-Crawler beginnt mit einer Liste von URLs und folgt dann den Hyperlinks von Webseite zu Webseite. Während er das Web durchforstet, ruft er den Inhalt jeder Seite ab, extrahiert alle relevanten Informationen und Daten und speichert die URL der Seite in einem Index. Dieser Vorgang wird vom Crawler für jede Seite in dem Index ständig wiederholt, wobei die Seite nach und nach aktualisiert wird. Auf diese Weise wird die Datenbank der Webseiten und der darin enthaltenen Informationen ständig erweitert.

Die Speicherung dieser Daten erfolgt dann an einem zentralen Speicherort/Index. Manchmal gibt es auch einen Zwischenspeicher, in dem neue Seiten gespeichert werden, bis sie an das Indexmodul weitergeleitet werden. Im Indexmodul werden die Seiten den gefundenen Begriffen oder Schlüsselwörtern zugeordnet und in dem Index aufgenommen, so dass das System schnell und effizient relevante Dokumente auf der Grundlage einer Benutzeranfrage abrufen kann. Dieser Prozess der Beschaffung, Speicherung und Verarbeitung der oben beschriebenen Daten findet ständig statt, auch wenn keine Abfragen durch einen Nutzer gestellt werden. Möchte ein Nutzer dann eine Suchanfrage an den Index stellen, wird die Anfrage zunächst in eine für das System verständliche Form gebracht. Dazu ist ein Abfragemodul (Query Module) erforderlich, das die Benutzeranfragen verarbeitet und interpretiert und dann die entsprechenden Daten aus dem System abruft.

Schließlich kommt nun der Teil, der für dieses Dokument wichtig ist: das Ranking-Modul. Das Ranking-Modul ordnet den vom Abfragemodul zurückgegebenen Daten eine Relevanzbewertung zu. Die Daten werden dann in der Reihenfolge ihrer Relevanz geordnet, wobei die relevantesten Dokumente in den Suchergebnissen zuerst erscheinen.

2.2. Herausforderungen von Web-Suchmaschinen

Die große Herausforderung für Web-Suchmaschinen ist die Größe des World Wide Web mit seiner enorm hohen Anzahl von Webseiten. Jeden Tag kommen etwa 252.000 neue Webseiten hinzu [31]. Die Identifizierung all dieser Webseiten erfordert daher eine Menge Ressourcen. Darüber hinaus können sich auch bereits gescannte Webseiten ändern, anpassen und neue Hyperlinks hinzufügen, wodurch wiederum neue Seiten vom Crawler-Modul erreicht werden können. Ein gutes Beispiel für diese beiden Herausforderungen ist Wikipedia. Allein auf Wikipedia gibt es bereits über 6,5 Millionen englische Artikel, von denen alle 2 Sekunden einer überarbeitet und aktualisiert wird. Darüber hinaus werden pro Tag etwa 560 neue Artikel hinzugefügt [62].

Zu dieser enormen Größe kommt noch das Problem hinzu, dass nicht alle URLs gescannt werden können, weil es viele unnötige Seiten gibt. Wenn Sie z. B. in einem Kalender auf "nächsten Monat" drücken, ist das eine neue URL, die eigentlich gescannt werden sollte. Wenn man sich jedoch all diese URLs betrachtet, gibt es praktisch eine unendliche Anzahl von URLs, die von den Web Crawlern gescannt werden müssten. Das wird natürlich nicht gemacht, weil es für den Nutzer der Suchmaschine keinen Zweck erfüllt und somit praktisch sinnlos ist.

Eine weitere Herausforderung, die Web-Suchmaschinen bewältigen müssen, ist die Tatsache, dass jeder eine Webseite betreiben kann. Eine Web-Suchmaschine muss also in der Lage sein, auf die Webseite einer Schule ebenso zuzugreifen wie auf die eines Museums oder auf die Webseite von etwas völlig Anderem.

Schlussendlich kommt es für den Nutzer hauptsächlich darauf an, wie der Ranking-Algorithmus eines Webbrowsers funktioniert.

3. Der PageRank-Algorithmus von Google

Laut Google liegt dem PageRank-Algorithmus die Idee zugrunde, dass eine Webseite mit größerer Wahrscheinlichkeit relevant und nützlich ist, wenn sie von anderen hochwertigen, relevanten Webseiten verlinkt wird. Mit anderen Worten: Der Algorithmus geht davon aus, dass die Qualität und Relevanz einer Webseite anhand der Seiten beurteilt werden kann, die auf sie verlinken. Der Algorithmus verwendet diese Informationen, um jeder Webseite einen "PageRank"-Wert zuzuweisen, wobei höhere Werte ein höheres Maß an Relevanz und Bedeutung anzeigen [16].

3.1. Anwendung

Der PageRank-Algorithmus wird von Google verwendet, um die Relevanz oder Wichtigkeit einer Seite zu bestimmen. Die genauen Details dieses Rankings können nicht genau bestimmt werden, da der genaue Algorithmus und die genaue Anzahl der Rankingfaktoren nicht bekannt sind. Es ist jedoch bekannt, dass der Suchalgorithmus von Google über 200 Faktoren berücksichtigt, die das Ranking einer Webseite beeinflussen, wobei der PageRank-Algorithmus nur ein Teil dieser Faktoren ist. Zu diesen gehören u.a. Faktoren der Domain, Faktoren auf Seitenebene, Faktoren auf Standortebene, Backlink Faktoren, die Interaktion des Nutzers, spezielle Algorithmus-Regeln, Marken-Signale, Webspam-Faktoren auf der Webseite und externe Webspam-Faktoren [12].

3.2. Weitere Anwendungsmöglichkeiten

Der Algorithmus kann jedoch auch in anderen Zusammenhängen angewendet werden, in denen es sinnvoll ist, Elemente nach ihrer Wichtigkeit oder Relevanz zu ordnen. Ein Beispiel für die Verwendung des PageRank-Algorithmus außerhalb von Suchmaschinen ist der Bereich der Analyse sozialer Netzwerke. In diesem Zusammenhang kann der Algorithmus verwendet werden, um die Bedeutung oder den Einfluss einzelner Benutzer innerhalb eines sozialen Netzwerks zu bestimmen. Durch die Analyse der Verbindungen zwischen den Nutzern und der Anzahl und Qualität der Links zwischen ihnen kann der PageRank-Algorithmus verwendet werden, um wichtige Einflussnehmer und wichtige Knotenpunkte innerhalb des Netzwerks zu ermitteln.

Ein weiteres Beispiel für den Einsatz des PageRank-Algorithmus ist die Verarbeitung natürlicher Sprache. In diesem Zusammenhang kann der Algorithmus eingesetzt werden, um die Bedeutung von Wörtern oder Sätzen in einem Dokument oder Textkorpus zu bewerten. Durch die Analyse der Häufigkeit und des gemeinsamen Auftretens von Wörtern und Phrasen kann der PageRank-Algorithmus dazu beitragen, die wichtigsten Themen und Konzepte innerhalb eines Textes zu identifizieren, wodurch das Verständnis und die Analyse des Inhalts erleichtert werden.

4. Algorithmus

Der Algorithmus ermittelt den PageRank (PR) einer Webseite, indem er die Anzahl der Links zu einer Webseite, die Qualität dieser Links und die Bedeutung der Webseiten, die auf die Webseite verlinken, berücksichtigt. Der PageRank-Algorithmus hilft nicht nur dabei, die Relevanz und Bedeutung von Webseiten zu bestimmen, sondern spielt auch eine Rolle bei der Bekämpfung von Spam und minderwertigen Webseiten. Indem er sich auf die Qualität und Relevanz von Links konzentriert, kann der Algorithmus dazu beitragen, Webseiten zu identifizieren und auszusortieren, die versuchen, ihr Ranking mit unethischen Mitteln zu manipulieren, z. B. durch den Kauf oder Handel von Links.

Da es viel schwieriger ist, auch die Qualität eines Links zu berücksichtigen, betrachte ich, wie oben erwähnt, nur eine modifizierte Version der Standardversion des Algorithmus, die jeden Link als gleichwertig betrachtet.

4.1. Aufbau des Algorithmus

Um das Internet effektiver beurteilen zu können, wird das World Wide Web als ein globales Netz von Milliarden miteinander verbundener Seiten betrachtet. Ein solches Netzwerk kann auch als Graph verstanden werden.

4.1.1. Das Internet als Graph

Das Internet kann als Graph modelliert werden, d. h. als eine mathematische Struktur, die aus einer Reihe von Knoten besteht, die die einzelnen Webseiten im Internet darstellen, und einer Reihe von Kanten oder Links, die die Verbindungen oder Pfade zwischen diesen Knoten darstellen. Durch die Darstellung des Internets als Graph ist es möglich, das Netz und seine Struktur auf eine Weise zu visualisieren und zu analysieren, die für die Optimierung seiner Leistung nützlich ist. Da Links nur in eine Richtung funktionieren, d. h. von einer Webseite zur anderen, ist das Internet ein gerichteter Graph G.

Definition 4.1: Bei einem gerichteten Graph $G = (V,\ E)$ *, sei* V *die endliche Menge von Knoten (Webseiten) und* E *die Menge gerichteter Kanten (Hyperlinks) zwischen den Seiten* [vgl. 1]. *Wobei:*

- $N :=$ *Gesamtanzahl der Knoten im Graphen* G
- *Die Menge aller Nachfolger eines Knotens v wird mit* $S_G(v) = \{u|(v,u) \in E\}$ *bezeichnet und die Menge aller Vorgänger mit* $P_G(v) = \{u|(u,v) \in E\}$ [1]
- $N_v = |S_G(v)|$ *sei die Anzahl der ausgehenden Verbindungen von* v

Eine Eigenschaft dieses Graphen ergibt sich aus der Definition des Internets, nämlich, dass jeder Knoten auf irgendeine Weise erreichbar ist. Denn wenn ein Knoten nicht erreichbar ist, ist er nicht im Internet vorhanden. Außerdem kann das Internet als dünnbesetzten Graphen betrachtet werden, da für etwa 75% der Knoten v die Anzahl der ausgehenden Verbindungen N_v bei höchstens zwei beträgt [42].

4.1.2. Definition PageRank

Der PageRank-Algorithmus gibt für jede vom Algorithmus bewertete Webseite eine Punktzahl aus. Dieser Wert, der oft als "PageRank" oder "PR"-Wert bezeichnet wird, ist ein numerischer Wert, der die vom Algorithmus ermittelte Bedeutung und Relevanz einer Webseite angibt. Dieser Wert ist in der Regel eine Wahrscheinlichkeitsverteilung, die angibt, wie wahrscheinlich es ist, dass eine Person auf diese Seite klicken wird.

Die ursprüngliche Definition des PageRank-Algorithmus von Brin und Page lautet [9]:

"ACADEMIC CITATION LITERATURE HAS BEEN APPLIED TO THE WEB, LARGELY BY COUNTING CITATIONS OR BACKLINKS TO A GIVEN PAGE. THIS GIVES SOME APPROXIMATION OF A PAGE'S IMPORTANCE OR QUALITY. PAGERANK EXTENDS THIS IDEA BY NOT COUNTING LINKS FROM ALL PAGES EQUALLY, AND BY NORMALIZING BY THE NUMBER OF LINKS ON A PAGE. PAGERANK IS DEFINED AS FOLLOWS:

WE ASSUME PAGE A HAS PAGES T1...TN WHICH POINT TO IT (I.E., ARE CITATIONS). THE PARAMETER D IS A DAMPING FACTOR WHICH CAN BE SET BETWEEN 0 AND 1. WE USUALLY SET D TO 0.85. THERE ARE MORE DETAILS ABOUT D IN THE NEXT SECTION. ALSO C(A) IS DEFINED AS THE NUMBER OF LINKS GOING OUT OF PAGE A. THE PAGERANK OF A PAGE A IS GIVEN AS FOLLOWS:

PR(A) = (1-D) + D (PR(T1)/C(T1) + ... + PR(TN)/C(TN))

NOTE THAT THE PAGERANKS FORM A PROBABILITY DISTRIBUTION OVER WEB PAGES, SO THE SUM OF ALL WEB PAGES' PAGERANKS WILL BE ONE."

Aus dieser Definition lässt sich Folgendes ableiten:

Definition 4.2: Um ein Gewicht PR_{p_i} einer Seite p_i zu berechnen, werden die Gewichte PR_{p_j} der Seiten p_j benötigt, die auf p_i verweisen. Wenn p_j auf N_{p_j} Seiten verweist, wird das Gewicht PR_{p_j} auf diese Seiten verteilt. Der PR_{p_i} für jede Seite p_i wird durch diese rekursive Formel ausgedrückt:

$$PR_{p_i} = \frac{1-d}{N} + d \sum_{p_j \in P_G(p_i)} \frac{PR_{p_j}}{N_{p_j}} \quad \textit{[vgl. 77]}$$

Die rekursive Formel wird auch ohne den Normalisierungsfaktor $\frac{1}{N}$ angegeben, da dieser in der Originalarbeit von Brin und Page ebenfalls nicht vorhanden ist [9].

Dabei ist die Menge der Seiten $V = \{p_1, p_2, \dots p_N\}$, N ist die Anzahl der Knoten des Graphen G und $d \in [0; 1]$ ist der Dämpfungsfaktor. Der Dämpfungsfaktor wird verwendet, um das Gewicht von Links zu reduzieren, indem ein Teil des Linkgewichts auf alle Webseiten im Index verteilt wird. Dadurch wird verhindert, dass eine Webseite aufgrund einer unnatürlich hohen Anzahl von Links von anderen Webseiten zu hoch eingestuft wird.

Der Dämpfungsfaktor wird in der Regel auf einen Wert zwischen 0 und 1 gesetzt, wobei ein niedrigerer Wert eine stärkere Abschwächung des Linkgewichts bedeutet. Durch die Einbeziehung des Dämpfungsfaktors ist der Algorithmus in der Lage, die Relevanz und Bedeutung einer Webseite auf der Grundlage der Qualität und Relevanz der Links zu ihr genauer wiederzugeben. In den meisten Fällen beträgt dieser Faktor 0.85, wie von Brin und Page vorgeschlagen [9].

Der PR_{p_i} einer Seite p_i ist daher gleich 1 minus dem Dämpfungsfaktor, plus dem PR_{p_j} jeder Seite p_j, die auf p_i verweist, dividiert durch die Anzahl der von p_j ausgehenden Links, reduziert um den Dämpfungsfaktor d.

4.2. Surfer Modell

Eine intuitive Variante sich die Funktionsweise des Algorithmus vorzustellen ist, das Surfer Modell. Beim Surfer-Modell handelt es sich um ein hypothetisches Modell des Nutzerverhaltens, das zur Anpassung der Bedeutung bestimmter Faktoren bei der PageRank-Berechnung verwendet wird.

4.2.1. Random Surfer-Modell

Die Grundidee hinter dem Random-Surfer-Modell ist, dass es das Verhalten eines Nutzers darstellt, der zufällig auf Links im Internet klickt, ohne ein bestimmtes Ziel oder einen bestimmten Ort im Sinn zu haben. Dieser hypothetische Nutzer wird "Random Surfer" genannt, weil er zufällig im Internet surft und den Links folgt, auf die er stößt, ohne eine bestimmte Richtung oder Absicht zu haben. Ein solches mathematische Modell wird als "Random Walk" bezeichnet.

Das Random-Surfer-Modell wird verwendet, um die Bedeutung bestimmter Faktoren bei der Berechnung des PageRank-Wertes einer Webseite anzupassen, z. B. den Dämpfungsfaktor d. Der Algorithmus berücksichtigt zum Beispiel die Anzahl der Links zu einer Webseite sowie die Qualität und Relevanz dieser Links, da Webseiten mit einem höheren PR für den zufälligen Surfer wichtiger sind und daher eher angeklickt werden. Ohne das Random-Surfer-Modell könnte eine Webseite mit einer großen Anzahl von minderwertigen oder irrelevanten Links immer noch einen hohen PageRank-Wert haben.

4.2.2. Rational Surfer Modell

Das Modell des rationalen Surfers ist eine Erweiterung des Modells des zufälligen Surfers, bei dem im Gegensatz zum Modell des zufälligen Surfers das Verhalten eines Nutzers dargestellt wird, der auf Links mit einem bestimmten Ziel oder Standort im Kopf klickt. Dieser hypothetische Nutzer wird als "rationaler Surfer" bezeichnet, weil er bewusste und rationale Entscheidungen darüber trifft, welchen Links er folgt, um sein gewünschtes Ziel zu erreichen.

4.3. Berechnung des PageRank

Bei der Formel des PageRank-Algorithmus (Definition 4.2) tritt jedoch das Problem auf, dass die Werte PR_{p_j} nicht bekannt sind. Um dieses Problem zu beheben geht man davon aus, dass jede Seite p_j zu Beginn den gleichen PR-Wert, nämlich $\frac{1}{N}$, hat.

4.3.1. Iterative Berechnung

Dieser Ansatz führt zu dieser neuen Definition:

Definition 4.3: Um die Formel (Definition 4.2) dem Iterativen Prozess anzupassen wird nun noch einen Parameter $k \in \mathbb{N}$ hinzugefügt, der die Iterationsschritte zählt. Bei $k = 0$ ist der PR jeder Seite $\frac{1}{N}$:

$$PR_{p_i}^{k+1} = \begin{cases} \frac{1}{N} & , wenn\ k = 0 \\ \frac{1-\mathrm{d}}{\mathrm{N}} + \mathrm{d} \sum_{p_j \in P_G(p_i)} \frac{PR_{p_j}^{k}}{N_{p_j}} & , sonst \end{cases}$$ [vgl. 36]

Die Iteration wird sooft durchgeführt, bis der *PR* hoffentlich in einen finalen Wert konvergiert.

4.3.2. Matrizenberechnung

Für eine effektive und einfache Berechnung des PR eines Graphen wird nun eine Hyperlink-Matrix H eingeführt, um den Graphen darzustellen. Diese Hyperlink-Matrix stellt die Verbindungen zwischen den Seiten p_i und p_j dar.

Definition 4.4: Die $N \times N$-Matrix H ist wie folgt definiert:

$$[\mathrm{H}]_{i,j} = \begin{cases} \frac{1}{N_{p_i}}, & (p_i, p_j) \in E \\ 0, & sonst \end{cases}$$ [vgl. 1]

Aufbau der Matrix H:

$$\begin{pmatrix} \frac{1}{N_{p_1}}, & (p_1, p_1) & \cdots & \frac{1}{N_{p_j}}, & (p_1, p_i) \\ & \vdots & \ddots & & \vdots \\ \frac{1}{N_{p_i}}, & (p_i, p_1) & \cdots & \frac{1}{N_{p_i}}, & (p_i, p_i) \end{pmatrix}$$

Diese Matrix ähnelt der Adjazenzmatrix des Graphen G, aber die Matrix H hat den entscheidenden Unterschied, dass jeder Zeilenwert einer Seite p_i in der Matrix das Verhältnis zwischen dem Link, der auf p_j landet, und der Gesamtanzahl an Links auf der Seite p_i ist, wobei Null angibt, dass es keinen direkten Weg von p_i zu p_j gibt. Nach dieser Definition der Hyperlink-Matrix handelt es sich um eine zeilennormierte Matrix, d. h. eine Matrix, in der jede Zeile durch die Anzahl der von der entsprechenden Webseite ausgehenden Links geteilt wird. Dies gibt an, wie viel "Gewicht" jeder Link hat.

Mit Hilfe der Potenzmethode [67] kann nun der PR jeder Seite berechnet werden. Die Potenzmethode ist eine Methode zur Berechnung des Haupteigenvektors einer Matrix. Dieser Vektor $\pi^{(k)}$ wird als PageRank-Vektor bezeichnet, da er alle PR-Werte der Webseiten darstellt.

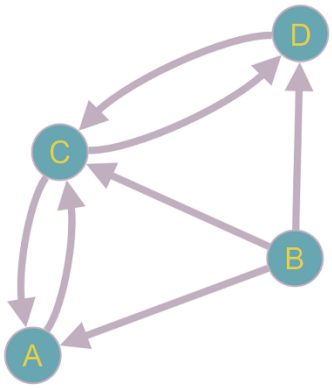
Definition 4.5: Der Vektor $\pi^{(k)}$ gibt ein $1 \times n$-Vektor der k-ten Iteration an. Dieser Vektor ist in der Nullten Iteration definiert als: $\pi^{(0)} = \frac{1}{N} e$ (mit e als $1 \times n$ Vektor mit nur Einsen, da zu Beginn jede Seite die gleiche Wahrscheinlichkeit hat, besucht zu werden) [36, Seite 44]

Definition 4.6: Mit dieser Definition vom PageRank Vektor, lässt sich nun die nächste Iteration des Vektors wie folgt berechnen:

$\pi^{(k+1)} = \pi^{(k)} H$ [ebd., *Seite 44]*

Die Vorteile einer solchen Berechnung sind, dass H eine sehr dünn besetze Matrix ist, d.h. die meisten Elemente dieser Matrix sind Nullen, da nur wenige Seiten auf andere Seiten verweisen, wodurch die Matrix-Multiplikationen einen minimalen Rechenaufwand aufweist
[ebd., Seite 44].

Beispiel für die Berechnung von PR-Werten mit der Matrixberechnung für einen Graphen G_4:



$N = 4$

$d = 0.85$

$$H = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0.33 & 0 & 0.33 & 0.33 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Ranking nach 3 Iterationen.
PR_A = 1/4	0.21	0.28	0.22	2
PR_B = 1/4	0.038	0.038	0.038	4
PR_C = 1/4	0.53	0.41	0.52	1
PR_D = 1/4	0.21	0.28	0.22	2

4.3.3. Probleme der iterativen Berechnung

Allerdings gibt es bei dieser Berechnung noch Probleme. Zum einen lässt sich nicht feststellen, wie lange der Berechnungsprozess dauert, bis die PR-Werte konvergieren, falls dies überhaupt geschieht. Zum anderen muss man sich mit den Seiten beschäftigen, die als Schleifen oder Senken fungieren. Eine Senke ist eine Seite, die keine weiteren Links ausgibt und somit den PR-Wert " schluckt". Das Problem bei Schleifen ist, dass dieser Prozess nie konvergiert und somit unendlich lange andauert [ebd., Seite 44].

4.3.4. Google Matrix

Um diese Probleme zu lösen, wird die Google-Matrix eingeführt. Die Google Matrix ist die Endgültig verwendete Matrix des PageRank-Algorithmus. Die Matrix wird verwendet, um den PR für alle Seiten gleichzeitig zu berechnen.

Definition 4.7: Um die Problem zu beheben, wird eine weitere Matrix S eingeführt. Bei dieser Matrix handelt es sich um eine Übergangsmatrix, da nun auch Seiten beachtet werden, die keine weiterführenden Links haben. Damit dies gelingt addiert man zu der Matrix H, $\frac{1}{N}e$ falls die Seite p_i eine solche Seite ist.

$$S = H + a(\frac{1}{N}e) \quad \text{mit } [a]_i = \begin{cases} 1, & N_{p_i} = 0 \\ 0, & sonst \end{cases}$$ [vgl. 36],

wobei es sich bei a um einen $n \times 1$-Vektor handelt.

Diese Veränderung führt dann dazu, dass die Matrix H von einer nicht stochastischen Matrix zu der Matrix S wird, welche stochastisch ist. Eine solche Matrix ist nun eine Übergangswahrscheinlichkeitsmatrix für eine Markow-Kette.

Eine Markow-Kette ist ein mathematisches Modell, das verwendet wird, um zufällige Prozesse zu beschreiben, die sich in verschiedenen Zuständen befinden können. Im Zusammenhang mit dem PageRank-Algorithmus wird eine Markow-Kette verwendet, um das Verhalten von Nutzern im Netzwerk von Links zwischen Webseiten zu beschreiben. Die Zustände in der Markow-Kette entsprechen dabei den verschiedenen Webseiten im Netzwerk, und die Übergänge zwischen den Zuständen werden durch die Links zwischen den Webseiten definiert. Die Verwendung einer Markow-Kette erleichtert die Berechnung der Wahrscheinlichkeit, dass ein Nutzer von einer Webseite zu einer anderen navigiert.

Nun ist zu bedenken, dass die Nutzer nicht nur wahllos von Link zu Link gehen, sondern manchmal auch von einer neuen Seite starten (dies würde der Eingabe einer neuen URL in der Suchleiste entsprechen).

Definition 4.8: Wenn man dies berücksichtigt, kommt man schließlich auf die Google-Matrix G. Mathematisch wird dies durch diese Formel ausgedrückt:

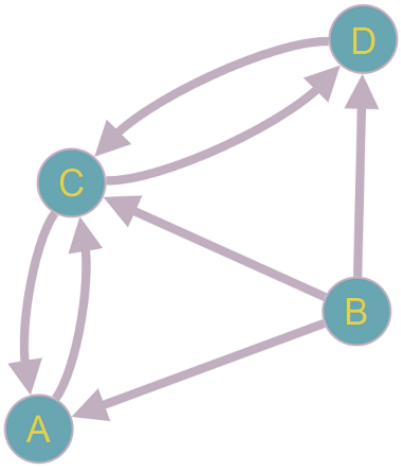
$$G = dS + \frac{(1-d)}{N}E_G$$ [vgl. 59]

Hier stellt $\frac{1}{N}E_G$ die „Teleportations-Matrix“ dar, die davon ausgeht, das sich der Nutzer von jeder Seite p_i zu einer anderen Seite p_j gelangen kann mithilfe einer URL (Die Teleportations-matrix ist eine $N \times N$-Matrix mit nur Einsen).

Da die Google-Matrix G quasi immer noch die Hyperlink-Matrix H ist, lässt sich nun der PageRank der Matrix mit einer angepassten Formel aus Definition 4.6 iterativ berechnen:

$\pi^{(k+1)} = \pi^{(k)} G$ [36, Seite 49]

Beispiel für die Berechnung von PR-Werten mit der Google-Matrix für einen Graphen G_4:



$N = 4$

$d = 0.85$

$$H = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0.33 & 0 & 0.33 & 0.33 \\ 0.5 & 0 & 0 & 0.5 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

$$G = \begin{pmatrix} 0.038 & 0.038 & 0.89 & 0.038 \\ 0.32 & 0.038 & 0.32 & 0.32 \\ 0.46 & 0.038 & 0.038 & 0.46 \\ 0.038 & 0.038 & 0.89 & 0.038 \end{pmatrix}$$

$\pi^{(k)} = (0.25 \quad 0.038 \quad 0.47 \quad 0.25)$

4.3.5. Berechnung des PageRank Vektors

Mit dieser Formel lässt sich nun $\pi^{(k+1)}$ berechnen:

$$\pi^{(k+1)} = \pi^{(k)} G \qquad | \; G = dS + \frac{(1-d)}{N} E_G$$

$$\pi^{(k+1)} = d\pi^{(k)} S + \frac{1-d}{N} \pi^{(k)} E_G \qquad | \; S = H + a(\frac{1}{N} e)$$

Definition 4.9: Man formt nun die Berechnung von $\pi^{(k+1)}$ zu dieser Gleichung um:

$$\pi^{(k+1)} = d\pi^{(k)} H + \left(d\pi^{(k)} a + (1-d)\right) \frac{e}{N}$$ [vgl. 36, Seite 49]

Diese Gleichung wird zur Berechnung der nächsten Annäherung an den Eigenwert bei der Potenzmethode verwendet. Die Gleichung wird in dieser Form angegeben, weil die Google-Matrix eine dichte Matrix ist, d. h. eine Matrix, in der die meisten Elemente ungleich Null sind. Mit anderen Worten: Eine dichte Matrix hat eine relativ hohe Anzahl von Nicht-Null-Elementen im Vergleich zur Gesamtzahl der Elemente in der Matrix. H hingegen ist eine

dünnbesetzte Matrix, die es ermöglicht, durch verschiedene Komprimierungsverfahren Speicherplatz zu sparen und so den PR-Vektor effizienter zu berechnen.

5. Ranking der Suchergebnisse

Führt man diese Berechnung nun für den PageRank-Vektor durch, erhält man einen PageRank-Vektor π, der die Bedeutung jeder Webseite im Index berechnet.

Um den PageRank-Vektor in einen Rang in den Suchergebnissen umzuwandeln, ermittelt die Suchmaschine zunächst die Schlüsselwörter, die ein Nutzer in seine Suchanfrage eingegeben hat. Dann durchsucht sie ihren Index von Webseiten, um diejenigen zu finden, die für die Schlüsselwörter relevant sind. Für jede relevante Webseite berechnet die Suchmaschine ihren PageRank-Wert anhand des PageRank-Vektors. Die Suchmaschine sortiert dann die relevanten Webseiten nach ihrem PageRank-Wert und präsentiert sie dem Nutzer in der Reihenfolge ihrer Wichtigkeit, wobei die relevantesten Webseiten ganz oben in den Suchergebnissen erscheinen

6. Implementierung und Deutung am eigenen Beispiel: Wikipedia

Nachdem die Berechnung des PageRank-Algorithmus und seine Rolle bei der Bewertung von Webseiten erläutert wurde, wird nun die Implementierung der Programmierung behandelt. Durch die Implementierung des PageRank-Algorithmus kann die Wichtigkeit jeder Webseite innerhalb des Indexes berechnet werden und diese Information kann verwendet werden, um relevante Suchergebnisse zu sortieren und dem Benutzer zu präsentieren. Zu diesem Zweck wird Wikipedia als Datenquelle für den Index verwendet. Zunächst muss jedoch der PageRank-Algorithmus implementiert werden [A2].

Um den PageRank-Algorithmus zu implementieren, wird eine Hyperlink-Matrix H, ein PR-Ausgangsvektor, ein Dämpfungsfaktor d und die Anzahl der Knoten in der Matrix benötigt. Darüber hinaus werden der a-Vektor und die e-Matrix aus der Hyperlink-Matrix berechnet, um anschließend die Formel aus Definition 4.9 zur Berechnung der nächsten Iteration des PR-Vektors zu implementieren.

Eine mögliche Implementierung könnte daher wie folgt aussehen [A3 matrix.py (angepasst)]:

```
def calculate_next_pr_vector(array, pr_vector, damping_factor, numOfV):
    H = array
    pi_k = pr_vector
    d = damping_factor
    N = numOfV
    a = compute_a_vector()
    e = compute_e_matrix()

    d_pi_k_H = np.dot(d*pi_k, H)
    d_pi_k_a = np.dot(d*pi_k, a)
    partial_calculation = np.dot(d_pi_k_a + (1-d), e/N)
    new_pr_vector = np.add(d_pi_k_H, partial_calculation)
    return new_pr_vector
```

Nachdem der PageRank-Algorithmus nun implementiert ist, können seine Genauigkeit und Leistung anhand echter Daten getestet werden. Zu diesem Zweck wird, wie bereits erwähnt, Wikipedia als Datenquelle für diese Untersuchung verwendet. Für diese Wahl gibt es mehrere Gründe. Erstens ist Wikipedia eine weit verbreitete und angesehene Online-Enzyklopädie mit einer großen und vielfältigen Sammlung von Artikeln zu einer Vielzahl von Themen. Dies bedeutet, dass ein großer und vielfältiger Datensatz zur Verfügung steht, der repräsentativ für die Arten von Webseiten und Informationen ist, die in der Praxis häufig verwendet werden. Zweitens ist Wikipedia eine Open-Source-Plattform, was bedeutet, dass der Inhalt für jedermann zur Nutzung und Erforschung zur Verfügung steht. Dies macht sie zu einer idealen Ressource für die akademische Forschung, da es leicht ist, auf die Daten zuzugreifen und sie zu analysieren, ohne Kosten zu verursachen oder Genehmigungen einholen zu müssen. Dies macht Wikipedia zu einem nützlichen Maßstab für die Bewertung der Leistung des PageRank-Algorithmus, da die Ergebnisse des Algorithmus mit der realen Relevanz und Glaubwürdigkeit von Wikipedia-Artikeln verglichen werden können.

Um jedoch die Implementierung des PageRank-Algorithmus auf einen Wikipedia-Datensatz anzuwenden, muss dieser Datensatz zunächst gekürzt und nur die wichtigsten Daten, d.h. Verweise auf andere Wikipedia-Artikel, extrahiert werden. Dann wird aus diesen Daten eine Matrix erstellt, mit der die Berechnung erneut durchgeführt werden kann. Dabei ist zu beachten, dass es 5.292.520 deutsche Wikipedia-Artikel gibt. Um all diese in der Matrix H zu speichern, wird eine CRS-Matrix (Compressed Row Storage-Matrix) verwendet.

Aufbau deiner Seite im XML-Dump

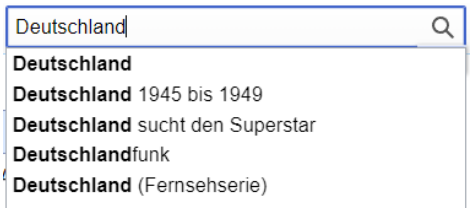
Nachdem nun der PR-Vektor für Wikipedia berechnet wurde, braucht es nur noch eine Suchmaschine, die die Anfrage verarbeitet und dann die relevanten Seiten für unsere Anfrage ausgibt, sortiert nach dem PR-Vektor. Um die Ausgabe so realitätsnah wie möglich zu halten, soll zunächst die Seite angezeigt werden, die genau dem Titel einer Seite entspricht. Dann wird jeder Wikipedia-Artikel daraufhin überprüft, ob der Suchbegriff im Titel vorkommt, und dann wird jeder Artikel nach dem PR-Wert sortiert, wobei der Artikel mit dem größeren PR-Wert am höchsten angezeigt wird.

Nun können einige Suchen durchgeführt und die Ergebnisse mit der Glaubwürdigkeit und Relevanz von Wikipedia-Artikeln verglichen werden, um die Leistung und Nützlichkeit des implementierten PageRank-Algorithmus zu überprüfen. Dabei wird keine erweiterte Suche, weder in der Implementierung noch in Wikipedia, verwendet.

Suchanfragen 1: „Deutschland“

Die ersten fünf Ergebnisse des implementierten PageRank-Algorithmus sind: Deutschland, Deutschlandfunk, Deutschlandradio, Deutschlandfunk Kultur, Deutschlandradio Kultur

Auf Wikipedia lauten die ersten fünf Ergebnisse:

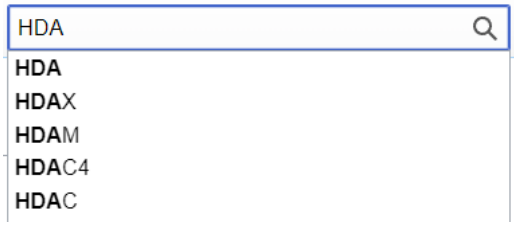


Zwei der fünf Ergebnisse sind bei beiden Rankings identisch. Man kann also sagen, dass in diesem Beispiel die Umsetzung des PageRank-Algorithmus teils effektiv ist wie die von Wikipedia. Kultur und Rundfunkanstalten sind jedoch überrepräsentiert. Allerdings gibt es noch das Problem, dass Deutschlandradio Kultur auf genau denselben Wikipedia-Artikel verweist wie Deutschlandfunk Kultur. Abgesehen von diesem Problem, nicht unterscheiden zu können, ob ein Artikel auf einen anderen Artikel verweist, kann man sagen, dass dieses Beispiel für die Implementierung eines solchen PageRank-Algorithmus teils spricht.

Suchanfragen 2: „HDA“

Die ersten fünf Ergebnisse des implementierten PageRank-Algorithmus sind: HDA, HDAX, HDAC4, HDAC, HDAV

Bei Wikipedia sind die ersten fünf Ergebnisse:



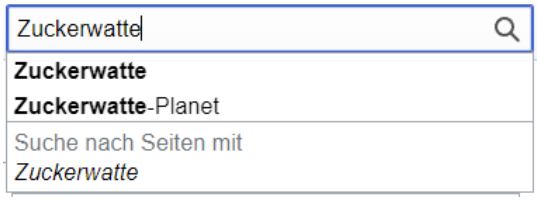
In diesem sehr spezifischen Beispiel stimmt nur eines der Ergebnisse nicht überein. Daraus lässt sich schließen, dass der Algorithmus bei sehr kurzen, spezifischen Suchanfragen effektiver arbeitet als bei längeren, komplexeren Abfragen.

Suchanfragen 3: „Zuckerwatte“

Als letztes Beispiel wird ein bestimmtes Wort genommen, das gleichzeitig nicht so viele Artikel haben soll wie "Deutschland".

Die ersten fünf Ergebnisse des implementierten PageRank-Algorithmus sind: Zuckerwatte, Zuckerwatte-Planet, Zuckerwattemaschine, Zuckerwattebecher

Bei Wikipedia sind die ersten fünf Ergebnisse:



Die integrierte Suchleiste von Wikipedia liefert nur zwei Ergebnisse. Diese beiden stimmen jedoch genau mit dem implementierten PageRank-Algorithmus überein. Daraus lässt sich ableiten, dass bestimmte längere und eindeutigere Suchanfragen mit dem implementierten Algorithmus mehr Ergebnisse liefern als mit Wikipedia.

Im Allgemeinen lässt sich aus diesen Beispielen ableiten, dass der PageRank-Algorithmus effektiv implementiert wurde und daher repräsentativ für eine reale Implementierung des Algorithmus ist. Manchmal ist der PageRank-Algorithmus sogar effektiver und liefert mehr Ergebnisse als der Wikipedia-Algorithmus. Dies liegt jedoch daran, dass der PageRank-Algorithmus jedes Mal jeden Artikel im Index durchläuft, was ihn viel langsamer als den Wikipedia-Algorithmus macht.

7. Geheime Erweiterung des PageRank-Algorithmus

Jetzt, da eine effektive Implementierung des PageRank-Algorithmus vorliegt, ist es möglich zu untersuchen, wie sich Erweiterungen auf die Suchergebnisse auswirken. So kann nun die Leitfrage: "Wie könnte eine geheime Erweiterung des PageRank-Algorithmus von Google aussehen?" beantworten werden.

Um über eine geheime Erweiterung des Algorithmus nachzudenken, muss man sich zunächst überlegen, welchen Grund Google für eine solche Erweiterung oder Anpassung seines Algorithmus haben könnte. Einerseits ist Google von seinen Nutzern abhängig, was bedeutet, dass eine solche Änderung oft auf dem Feedback der Nutzer beruht oder auf der Behebung von Problemen, die bei Tests und Analysen festgestellt wurden. So kann eine solche Erweiterung beispielsweise die Leistung verbessern oder zusätzliche Funktionen bereitstellen, um zu verhindern, dass Nutzer auf Alternativen ausweichen. Es ist auch möglich, dass Google Änderungen an seinem Algorithmus vornimmt, um der Konkurrenz voraus zu sein oder um sich einen Wettbewerbsvorteil auf dem Markt zu verschaffen. Im Allgemeinen ist es nicht ungewöhnlich, dass Unternehmen einige Aspekte ihrer Algorithmen geheim halten, um ihr geistiges Eigentum zu schützen oder andere daran zu hindern, ihre Technologie zu kopieren.

Um eine geheime Erweiterung zu finden, die mit den vorhandenen Mitteln umgesetzt werden kann, ist es notwendig zu prüfen, welche Daten im Wikipedia-Dump noch vorhanden sind. Für das bisherige Ranking der einzelnen Seiten wurden nur der Titel, die ID und Informationen aus dem Text übernommen. Was nun für eine Erweiterung Sinn machen würde, ist die Bewertung der Aktualität einer Seite. Hierfür können die "Timestamp"-Daten verwendet werden, die angeben, wann eine Seite zuletzt bearbeitet wurde. Dies ist besonders wichtig, damit aktuelle Ereignisse höher gerankt werden als solche, die nicht mehr so relevant sind.

Die Berücksichtigung der Aktualität einer Seite ist auch für den PageRank-Algorithmus von Google wichtig, da die bisher implementierte Standardversion des Algorithmus nur auf Links achtet, was bedeutet, dass neu erstellte Seiten, wie z. B. Zeitungsartikel, kaum von anderen

Seiten verlinkt werden, was ihren Rang senkt. Wenn nun auch die Aktualität berücksichtigt wird, kann sichergestellt werden, dass Seiten, die für ein bestimmtes Thema oder eine Suchanfrage besonders relevant sind, in den Suchergebnissen höher eingestuft werden. Dies könnte zu besseren Suchergebnissen für die Nutzer und zu einer insgesamt positiveren Nutzererfahrung führen.

Um diese Erweiterung durchzuführen, definieren wir zwei neue Vektoren λ und δ:

Definition 7.1: Der Vektor λ sei $n \times 1$-Vektor, der den Zeitpunkt angibt, zu dem die Seiten zuletzt bearbeitet wurden:

$[\lambda]_i = t(p_i), \quad \forall p_i \in V$, *wobei* $t(x)$ *den Zeitpunkt angibt, zu dem die Seite* x *zuletzt bearbeitet wurde*

Definition 7.2: Der Vektor δ sei $n \times 1$-Vektor, der das Verhältnis zwischen der Zeit der letzten Überarbeitung und der ältesten Seite, d.h. der ältesten Überarbeitung, wie folgt angibt:

$[\delta]_i = \frac{v_i - (\lambda)_{min}}{\sum_{j=1}^{N} v_j}, \quad \forall v_i \in \lambda$, *wobei* $(\lambda)_{min}$ *den Zeitpunkt der ältesten Überarbeitung angibt.*

Da der Vektor δ nun die Verhältnisse zwischen dem Zeitpunkt der letzten Revisionen und der Revision der ältesten Seite angibt, kann dieser Vektor mit der Formel aus Definition 4.9 verknüpft werden, was zu folgenden Anpassungen führt:

Definition 7.3: Wenn $\omega \in [0; 1]$ eine Zahl ist, die angibt, wie stark sich das Alter einer Seite den PR beeinflusst, ergibt sich aus Definition 4.9 die folgende Formel:

$$\pi^{(k+1)} = (1 - \omega)\left(d\pi^{(k)}H\right) + \omega\delta^T + (1 - \omega)\left(d\pi^{(k)}a + (1 - d)\right)\frac{e}{N}$$

Mit dieser Formel wird nun auch das Alter einer Seite bei der Berechnung des PR-Vektors berücksichtigt. Diese Anpassung des PageRank-Algorithmus dürfte nicht allzu viel ändern, da die meisten Wikipedia-Seiten ständig aktualisiert werden, um die neuesten Fakten wiederzugeben. Die Anpassung wirkt sich daher eher auf Wikipedia-Seiten aus, die seit langem nicht mehr aktualisiert wurden. Daher verschwinden die Seiten, die früher einen hohen PR-Wert hatten, aber nie aktualisiert wurden, aus den oberen Suchergebnissen.

Zum Beispiel ändert sich für den Suchbegriff "HDA" die Position von "HDA, HDAX, HDAC4, HDAC, **HDAV**" zu "HDA, HDAX, **HDAV**, HDAC4, HDAC". In diesem Fall hat die

Erweiterung die Seite "HDAV" von Position 5 auf Position 3 geändert. Solche Änderungen scheinen auf den ersten Blick nicht so bedeutsam zu sein, aber wenn man bedenkt, dass über 90 % der Seitenaufrufe bei Google auf die erste Seite entfallen, wird klar, dass selbst eine solche Änderung einen sehr starken Einfluss auf das Suchergebnis hat.

Bei der erweiterten Berechnung des PageRank-Algorithmus aus *Definition 7.3* müssen jedoch auch eine Reihe potenzieller Herausforderungen und Einschränkungen berücksichtigt werden. Ein potenzielles Problem besteht darin, dass es schwierig sein kann, die Aktualität einer Seite genau zu beurteilen, insbesondere wenn die Seite mehrere Themen abdeckt oder das Thema der Seite nicht klar definiert ist. Dies ist zwar kein Problem bei Wikipedia-Artikeln, wohl aber bei der Bewertung anderer Webseiten. Darüber hinaus kann die Einbeziehung der Aktualität in den PageRank-Algorithmus den Einsatz zusätzlicher Ressourcen und Rechenleistung erfordern, was die Gesamteffizienz des Algorithmus beeinträchtigen könnte.

Insgesamt ist es wichtig, die Vor- und Nachteile der Einbeziehung von der Aktualität in den PageRank-Algorithmus sorgfältig abzuwägen, bevor Änderungen am Algorithmus vorgenommen werden. Um ein gutes Verhältnis zwischen dem Standard-PR-Wert und der Aktualität einer Seite zu erhalten, kann der Vektor ω auf 0.25 gesetzt werden.

8. Modifizierungen

Im Laufe der Jahre gab es mehrere Variationen und Anpassungen des PageRank-Algorithmus und des allgemeinen Suchalgorithmus von Google. Einige relevante Beispiele sind:

8.1. Panda

Ziel dieser Anpassung des Ranking-Algorithmus war es, dass qualitativ minderwertige Webseiten einen noch niedrigeren PR-Wert und qualitativ hochwertige Webseiten einen noch höheren PR-Wert erhalten. Diese Anpassung wurde von Google "Panda" genannt und betraf 11,8 % der Suchanfragen. Laut Google sind qualitativ hochwertige Webseiten solche, die originelle Inhalte, ausführliche Berichte und durchdachte Analysen enthalten, während Webseiten von geringer Qualität solche sind, die den Nutzern wenig Mehrwert bieten, Inhalte von anderen Webseiten kopieren oder einfach nicht sehr nützlich sind. [52]

8.2. Google RankBrain

RankBrain ist eine künstliche Intelligenz, die 2015 von Google eingeführt wurde, um die Relevanz von Suchergebnissen zu verbessern [40]. Diese KI analysiert Nutzeranfragen, um den Kontext dahinter zu verstehen, und nutzt diese Informationen, um verschiedenen Faktoren je

nach Suchanfrage mehr oder weniger Bedeutung zuzuweisen, wodurch die Suchergebnisse für die Absicht des Nutzers relevanter werden. Ein großer Vorteil von RankBrain ist also, dass es sich um eine KI handelt, die ständig aus den Suchanfragen der Nutzer lernt.

8.3. Google Hummingbird

Der Google Hummingbird-Algorithmus ist die bedeutendste Anpassung des Suchalgorithmus, denn im Gegensatz zu "Panda" oder "RankBrain" hat er den gesamten Kernalgorithmus verändert. Im Hummingbird-Algorithmus ist der PageRank einer Seite nur noch einer von über 200 Faktoren [57]. Dieser Algorithmus wurde 2013 angekündigt und funktioniert ähnlich wie Googles RankBrain-Algorithmus, welcher zwei Jahre später angekündigt wurde, mit dem Unterschied, dass RankBrain dazu beitragen soll, die Relevanz von Suchergebnissen zu verbessern, während Hummingbird sich darauf konzentriert, die Fähigkeit der Suchmaschine zu verbessern, die Bedeutung und den Kontext von Anfragen als Ganzes zu verstehen.

9. Die Zukunft des PageRank-Algorithmus

Obwohl der "klassische" PageRank-Algorithmus in der Google-Suchmaschine kaum noch relevant ist, ist die Idee selbst immer noch von großer Bedeutung. Denn der Algorithmus ist immer noch das Grundprinzip vieler Suchalgorithmen im Web.

Es ist jedoch ein klarer Trend zu beobachten, nämlich, dass die Bedeutung von Links immer mehr an Wert verliert. Am Anfang basierte das gesamte Konzept von Google darauf. Später, mit der "Panda"-Anpassung, wurde auch die Qualität der Seiten in Betracht gezogen. Anschließend wurden auch die Suchanfragen der Nutzer ausgewertet und das Ranking in diesem Zusammenhang beeinflusst. Im Moment sind Links immer noch sehr wichtig, wenn auch nicht mehr so sehr wie damals. In Zukunft werden Links jedoch immer weniger Bedeutung haben, sagte John Mueller im "Search Off The Record"-Podcast auf die Frage nach der Bedeutung von Links. Allerdings sagte er auch, dass Links für Google nie völlig irrelevant werden, denn Google muss die Seiten im Internet irgendwie finden, und ohne Links ist das nicht möglich [49].

Die Zukunft von Ranking-Algorithmen wie dem PageRank-Algorithmus liegt meines Erachtens in der zunehmenden Einbindung von künstlicher Intelligenz (KI), wie es bei Googles RankBrain bereits der Fall ist. Die KI macht derzeit extrem große Fortschritte, was es auch sehr wahrscheinlich macht, dass KI eine immer wichtigere Rolle in Ranking-Algorithmen spielen wird. Eine mögliche Anwendung von KI in Ranking-Algorithmen ist der Einsatz von maschinellem Lernen, um die Absicht hinter einer Suchanfrage und den Kontext, in dem sie

gestellt wird, besser zu verstehen. Dies könnte es Suchmaschinen ermöglichen, eine Suchanfrage genauer mit relevanten und qualitativ hochwertigen Ergebnissen abzugleichen, wie z. B. Google Hummingbird. Es ist auch denkbar, dass KI zur Analyse des Nutzerverhaltens und der Beschäftigung mit den Suchergebnissen eingesetzt wird, z. B. zur Analyse der Verweildauer auf einer Webseite und der Anzahl der Klicks auf ein Ergebnis. Diese Informationen könnten genutzt werden, um die Ranking-Algorithmen weiter zu verfeinern und das Sucherlebnis für die Nutzer insgesamt zu verbessern.

Zusammenfassend lässt sich sagen, dass eine geheime Erweiterung wie die in dieser Arbeit vorgestellte nicht repräsentativ für die Zukunft des PageRank-Algorithmus ist. Dennoch zeigt diese Erweiterung, wie einfach es ist, die PR-Werte des Algorithmus zu ändern. Diese Änderung mag nicht so bedeutend erscheinen, aber gerade wegen der weiten Verbreitung von Google kann selbst eine solche Änderung starke Auswirkungen auf das Ranking vieler Tausend Seiten haben.

10. Anhänge

[A1] Handbuch zur Bedienung der Software

[A2] Dokumentation der Eigenleistung

[A3] Python Dateien

11. Literatur- und Quellenverzeichnis

3Blue1Brown. (2016, 6. August). *Vectors | Chapter 1, Essence of linear algebra* [Video]. YouTube. Abgerufen am 12. Oktober 2022, von https://www.youtube.com/watch?v=fNk_zzaMoSs

[1] Altman & Tennenholtz. (2005). *Ranking Systems: The PageRank Axioms*. Harvard University. Abgerufen am 18. Dezember 2022, von http://www.eecs.harvard.edu/cs286r/courses/fall11/papers/AT%2705.pdf

[2] Amine, A. (2021, 24. Dezember). *PageRank algorithm, fully explained - Towards Data Science*. Medium. Abgerufen am 18. Dezember 2022, von https://towardsdatascience.com/pagerank-algorithm-fully-explained-dc794184b4af

[3] *Antitrust: Commission fines Google €1.49 billion for abusive practices in online advertising*. (2019). European Commission. Abgerufen am 17. Dezember 2022, von https://ec.europa.eu/commission/presscorner/detail/en/IP_19_1770

[4] Armstrong, M. (2021, 6. August). *How Many Websites Are There?* Statista Infographics. Abgerufen am 18. Dezember 2022, von https://www.statista.com/chart/19058/number-of-websites-online/

[5] Axenovich. (2022). *Graph Theory*. Karlsruher Institut für Technologie. Abgerufen am 18. Dezember 2022, von https://www.math.kit.edu/iag6/lehre/graphtheory2019w/media/main.pdf

[6] Boldi, Santini & Vigna. (2009). *PageRank: Functional Dependencies*. Unimi. Abgerufen am 18. Dezember 2022, von https://vigna.di.unimi.it/ftp/papers/PageRankDependencies.pdf

[7] Borodin, Roberts, Rosenthal & Tsapapras. (2005). *Link Analysis Ranking: Algorithms, Theory, and Experiments*. Stanford University. Abgerufen am 18. Dezember 2022, von http://snap.stanford.edu/class/cs224w-readings/borodin05pagerank.pdf

[8] Brezinski & Redivo-Zaglia. (2006). *The PageRank Vector: Properties, Computation, Approximation, and Acceleration*. ResearchGate. Abgerufen am 18. Dezember 2022, von https://langvillea.people.cofc.edu/DISSECTION-LAB/ClarePageRankModule/Glossary.html

[9] Brin & Page. (1998). *The Anatomy of a Search Engine*. Stanford University. Abgerufen am 18. Dezember 2022, von http://infolab.stanford.edu/%7Ebackrub/google.html

[11] *Create Graph online and find shortest path or use other algorithm*. (o. D.). https://graphonline.ru/en/

[12] Dean. (2021, Oktober). *Google's 200 Ranking Factors: The Complete List (2022)*. Backlinko. Abgerufen am 18. Dezember 2022, von https://backlinko.com/google-ranking-factors

[13] Dean, J. A. (2004, 17. Juni). *Ranking documents based on user behavior and/or feature data*. Google. Abgerufen am 18. Dezember 2022, von https://patents.google.com/patent/US9305099B1/en

[14] Ding, He, Husbands, Zha & Simon. (2002). *PageRank, HITS and a Unified Framework for Link Analysis*. Society for Industrial and Applied Mathematics. Abgerufen am 18. Dezember 2022, von https://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.24

[15] *Extracting large-scale knowledge bases from the web*. (1999). IIT Bombay: Indian Institute of Technology Bombay. Abgerufen am 18. Dezember 2022, von https://www.cse.iitb.ac.in/~soumen/readings/papers/KumarRRT1999campfire.pdf

[16] *Facts about Google and Competition*. (2011). Internet Archive. Abgerufen am 18. Dezember 2022, von https://web.archive.org/web/20111104131332/https://www.google.com/competition/howgooglesearchworks.html

[17] Georgiadis, Italiano & Parotsidis. (2016). *2-Connectivity in Directed Graphs*. Schloss Dagstuhl. Abgerufen am 18. Dezember 2022, von https://drops.dagstuhl.de/opus/volltexte/2016/6345/pdf/LIPIcs-ESA-2016-1.pdf

[18] Gleich, Constantine, Flaxman & Gunawardana. (2010). *Tracking the Random Surfer: Empirically Measured Teleportation Parameters in PageRank*. ETH Zürich. Abgerufen am 18. Dezember 2022, von https://www.ra.ethz.ch/CDStore/www2010/www/p381.pdf

[19] GmbH, E. (2017). *Ingenieurkurse*. Einführung: Graphentheorie - Operations Research 1. Abgerufen am 18. Dezember 2022, von https://www.ingenieurkurse.de/unternehmensforschung/graphentheorie/einfuehrung-graphentheorie.html

[20] Google. (2008, 25. Juli). *We knew the web was big. . .* Official Google Blog. Abgerufen am 18. Dezember 2022, von https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

[21] *Google matrix*. (2011). Université Toulouse III. Abgerufen am 18. Dezember 2022, von https://www.quantware.ups-tlse.fr/dima/myrefs/myp15.pdf

[22] *Google matrix: fundamentals, applications and beyond*. (2018, 15. Oktober). Centre national de la recherche scientifique. Abgerufen am 18. Dezember 2022, von https://indico.math.cnrs.fr/event/3475/

[23] *GOOGLE RANKBRAIN: The Definitive Guide*. (2020). Backlinko. Abgerufen am 18. Dezember 2022, von https://backlinko.com/google-rankbrain-seo

[24] *Graph structure in the Web*. (2000). Stanford University. Abgerufen am 18. Dezember 2022, von https://snap.stanford.edu/class/cs224w-readings/broder00bowtie.pdf

[25] Hansell, S. (2007, 2. Juni). *Google Keeps Tweaking Its Search Engine*. The New York Times. Abgerufen am 18. Dezember 2022, von https://www.nytimes.com/2007/06/03/business/yourmoney/03google.html

[26] Haveliwala & Kamvar. (2003). *The Second Eigenvalue of the Google Matrix*. Stanford University. Abgerufen am 18. Dezember 2022, von https://nlp.stanford.edu/pubs/secondeigenvalue.pdf

[27] Haveliwala, Kamvar & Jeh. (2003). *An Analytical Comparison of Approaches to Personalizing PageRank.* Stanford University. Abgerufen am 18. Dezember 2022, von http://www-cs-students.stanford.edu/~taherh/papers/comparison.pdf

[28] Haynes, A. (2022, 19. April). *What Is Google PageRank? (+ Why Should You Care in 2022)*. Loganix. Abgerufen am 18. Dezember 2022, von https://loganix.com/google-pagerank/

[29] Holzmann, H., Anand & Khosla. (2019, 22. Oktober). *Estimating PageRank deviations in crawled graphs - Applied Network Science*. SpringerOpen. Abgerufen am 18. Dezember 2022, von https://appliednetsci.springeropen.com/articles/10.1007/s41109-019-0201-9

[30] *How Google retains more than 90% of market share*. (2018, 24. April). Business Insider. Abgerufen am 16. Dezember 2022, von https://www.businessinsider.com/how-google-retains-more-than-90-of-market-share-2018-4?international=true&r=US&IR=T

[31] Huss, N. (2022, 27. November). *How Many Websites Are There in the World?* Siteefy. Abgerufen am 18. Dezember 2022, von https://siteefy.com/how-many-websites-are-there/

[32] Jäger, G. (2009). *Mathematics for linguists*. Universität Tübingen. Abgerufen am 18. Dezember 2022, von http://www.sfs.uni-tuebingen.de/~gjaeger/lehre/ws0910/mathe/slides3.pdf

[33] *Justice Department Sues Monopolist Google For Violating Antitrust Laws*. (2020, 21. Oktober). OPA | Department of Justice. Abgerufen am 15. Dezember 2022, von https://www.justice.gov/opa/pr/justice-department-sues-monopolist-google-violating-antitrust-laws

[34] Kumar, Raghavan, Rajagopalan, Sivakumar, D., Tomkins & Upfal. (2010). *The Web as a graph*. Association for Computing Machinery. Abgerufen am 18. Dezember 2022, von https://dl.acm.org/doi/pdf/10.1145/335168.335170

[35] Langville & Meyer. (o. D.). *Deeper Inside PageRank*. Weierstrass Institute. Abgerufen am 18. Dezember 2022, von https://www.wias-berlin.de/people/koenig/www/DeeperInsidePRReprint.pdf

[36] Langville & Meyer. (2020). *Google's PageRank and Beyond: The Science of Search Engine Rankings*. Universität Bielefeld. Abgerufen am 18. Dezember 2022, von https://gi.cebitec.uni-bielefeld.de/_media/teaching/2019winter/alggr/langville_meyer_2006.pdf

[37] Levy, S. (2010, 22. Februar). *Exclusive: How Google's Algorithm Rules the Web*. WIRED. Abgerufen am 18. Dezember 2022, von https://www.wired.com/2010/02/ff-google-algorithm/

[38] Maslov & Redner. (2008). *Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks*. Journal of Neuroscience. Abgerufen am 18. Dezember 2022, von https://www.jneurosci.org/content/jneuro/28/44/11103.full.pdf

[39] Mathis & Julia. (2021). *PageRank-Algorithmus – FunFacts Wiki*. Universität Heidelberg. Abgerufen am 18. Dezember 2022, von https://funfacts.mathi.uni-heidelberg.de/index.php/PageRank-Algorithmus

[40] Moz. (2022, 1. Juni). *Understanding Google Rank Brain And How It Impacts SEO*. Abgerufen am 18. Dezember 2022, von https://moz.com/learn/seo/google-rankbrain

[41] *„Normalize“ values to sum 1 but keeping their weights*. (2013, 14. Januar). Mathematics Stack Exchange. Abgerufen am 18. Dezember 2022, von https://math.stackexchange.com/questions/278418/normalize-values-to-sum-1-but-keeping-their-weights

[42] *On Power-Law Relationships of the Internet Topology*. (1999). Association for Computing Machinery. Abgerufen am 18. Dezember 2022, von https://dl.acm.org/doi/epdf/10.1145/316194.316229

[43] Page, L. (1997, 10. Januar). *Method for node ranking in a linked database*. Google Patents. Abgerufen am 18. Dezember 2022, von https://patents.google.com/patent/US6285999

[44] Prystowsky & Gill. (2005). *Calculating Web Page Authority Using the PageRank Algorithm*. Harvard University. Abgerufen am 18. Dezember 2022, von https://courses.seas.harvard.edu/climate/eli/Courses/APM111/2006spring/supporting-material/01-linear-equations/Gill-Prystowsky-PageRank-explained.pdf

[45] Richardson & Domingos. (2001). *The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank*. University of Washington. Abgerufen am 18. Dezember 2022, von https://homes.cs.washington.edu/~pedrod/papers/nips01b.pdf

[46] Rogers. (2002). *The Google Pagerank Algorithm and How It Works*. Western Michigan University. Abgerufen am 18. Dezember 2022, von https://cs.wmich.edu/gupta/teaching/cs3310/lectureNotes_cs3310/Pagerank%20Explained%20Correctly%20with%20Examples_www.cs.princeton.edu_~chazelle_courses_BIB_pagerank.pdf

[47] Rossi & Gleich. (2012). *Dynamic PageRank Using Evolving Teleportation*. Ryan A. Rossi. Abgerufen am 18. Dezember 2022, von http://ryanrossi.com/pubs/rossi-gleich-dynamic-pagerank.pdf

[48] Sahu, Kothapalli & Banerjee. (2021). *Adjusting PageRank parameters and Comparing results*. arXiv. Abgerufen am 18. Dezember 2022, von https://arxiv.org/ftp/arxiv/papers/2108/2108.02997.pdf

[49] Schwartz, B. (2022, 4. November). *Google: Links Will Be Less Important As A Ranking Factor In The Future*. seroundtable.com. Abgerufen am 18. Dezember 2022, von https://www.seroundtable.com/google-links-less-important-ranking-34357.html

[50] Seo, H. (2022, 5. Dezember). *What is PageRank? PageRank Algorithm Definition and Analysis*. Holistic SEO. Abgerufen am 18. Dezember 2022, von https://www.holisticseo.digital/theoretical-seo/pagerank

[51] Sheldon. (2010, Februar). *MANIPULATION OF PAGERANK AND COLLECTIVE HIDDEN MARKOV MODELS*. UMass Amherst. Abgerufen am 18. Dezember 2022, von https://people.cs.umass.edu/~sheldon/papers/thesis.pdf

[52] Singhal. (2011). *Official Google Blog: Finding more high-quality sites in search*. Internet Archive. Abgerufen am 18. Dezember 2022, von https://web.archive.org/web/20111122053212/http://googleblog.blogspot.com/2011/02/finding-more-high-quality-sites-in.html

[53] Siu, T. (2019, 14. Juli). *Website Repository: What is it and why is it important? | KIMBO Design*. KIMBO Design. Abgerufen am 18. Dezember 2022, von https://www.kimbodesign.ca/website-repository-what-is-it-and-why-is-it-important/

[54] Slawski, B. (2021, 15. Juli). *Reasonable Surfer Model: How Link Value Differs Based on Link, Document Features and User Data*. SEO by the Sea. Abgerufen am 18. Dezember 2022, von https://www.seobythesea.com/2010/05/googles-reasonable-surfer-how-the-value-of-a-link-may-differ-based-upon-link-and-document-features-and-user-data/

[55] Strang. (2016). *Chapter 6 - Eigenvalues and Eigenvectors*. Massachusetts Institute of Technology. Abgerufen am 18. Dezember 2022, von https://math.mit.edu/~gs/linearalgebra/linearalgebra5_6-1.pdf

[56] Stumm, V., Junior. (2022, 3. Februar). *Link Analysis Algorithms Explained*. Zyte. Abgerufen am 18. Dezember 2022, von https://www.zyte.com/blog/link-analysis-algorithms-explained/

[57] Sullivan, D. (2022, 23. Februar). *FAQ: All About The New Google "Hummingbird" Algorithm*. Search Engine Land. Abgerufen am 18. Dezember 2022, von https://searchengineland.com/google-hummingbird-172816

[58] Sun, Xie, Zhang & Faloutsos. (2007). *Less is More: Compact Matrix Decomposition for Large Sparse Graphs*. Microsoft. Abgerufen am 18. Dezember 2022, von https://www.microsoft.com/en-us/research/wp-content/uploads/2007/04/sdm07.pdf

[59] *The Mathematics of PageRank*. (2020). College of Charleston. Abgerufen am 18. Dezember 2022, von https://langvillea.people.cofc.edu/DISSECTION-LAB/ClarePageRankModule/6_Mathematics.html

[60] *The Value of Google Result Positioning*. (2013, 7. Juli). Chitika Insights. Abgerufen am 11. November 2022, von https://research.chitika.com/wp-content/uploads/2022/02/chitikainsights-valueofgoogleresultspositioning.pdf

[61] *Web Algorithms for Information Retrieval:A Performance Comparative Study*. (2012). ResearchGate. Abgerufen am 18. Dezember 2022, von https://www.researchgate.net/publication/261484567_A_comparative_study_of_link_analysis_algorithms_for_information_retrieval

[62] Wikipedia contributors. (o. D.). *Wikipedia:Statistics - Wikipedia*. https://en.wikipedia.org/wiki/Wikipedia:Statistics

[63] Wikipedia contributors. (2022a, Januar 30). *Google matrix*. Wikipedia. https://en.wikipedia.org/wiki/Google_matrix

[64] Wikipedia contributors. (2022b, September 18). *Directed graph*. Wikipedia. https://en.wikipedia.org/wiki/Directed_graph

[65] Wikipedia contributors. (2022c, September 23). *Stochastic matrix*. Wikipedia. https://en.wikipedia.org/wiki/Stochastic_matrix

[66] Wikipedia contributors. (2022d, Oktober 17). *Strongly connected component*. Wikipedia. https://en.wikipedia.org/wiki/Strongly_connected_component

[67] Wikipedia contributors. (2022e, Oktober 18). *Power iteration*. Wikipedia. https://en.wikipedia.org/wiki/Power_iteration

[68] Wikipedia contributors. (2022f, November 1). *Adjacency matrix*. Wikipedia. https://en.wikipedia.org/wiki/Adjacency_matrix

[69] Wikipedia contributors. (2022g, November 8). *Connectivity (graph theory)*. Wikipedia. https://en.wikipedia.org/wiki/Connectivity_(graph_theory)

[70] Wikipedia contributors. (2022h, November 18). *Subset*. Wikipedia. https://en.wikipedia.org/wiki/Subset

[71] Wikipedia contributors. (2022i, November 28). *Web crawler*. Wikipedia. https://en.wikipedia.org/wiki/Web_crawler

[72] Wikipedia contributors. (2022j, Dezember 1). *Random walk*. Wikipedia. https://en.wikipedia.org/wiki/Random_walk

[73] Wikipedia contributors. (2022k, Dezember 9). *Markov chain*. Wikipedia. https://en.wikipedia.org/wiki/Markov_chain

[74] Wikipedia contributors. (2022l, Dezember 10). *Eigenvalues and eigenvectors*. Wikipedia. https://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors

[75] Wikipedia contributors. (2022m, Dezember 15). *Information technology*. Wikipedia. https://en.wikipedia.org/wiki/Information_technology

[76] Wikipedia contributors. (2022n, Dezember 15). *Sparse matrix*. Wikipedia. https://en.wikipedia.org/wiki/Sparse_matrix

[77] Wikipedia contributors. (2022o, Dezember 16). *PageRank*. Wikipedia. Abgerufen am 18. Dezember 2022, von https://en.wikipedia.org/wiki/PageRank

[78] Wikipedia contributors. (2022p, Dezember 17). *Graph theory*. Wikipedia. https://en.wikipedia.org/wiki/Graph_theory

[79] Wikipedia contributors. (2022q, Dezember 17). *Network topology*. Wikipedia. https://en.wikipedia.org/wiki/Network_topology

[80] Wikipedia-Autoren. (2001, 10. Dezember). *Suchmaschine*. https://de.wikipedia.org/wiki/Suchmaschine

[81] Wikipedia-Autoren. (2002, 21. November). *Eigenwertproblem*. https://de.wikipedia.org/wiki/Eigenwertproblem

[82] Wikipedia-Autoren. (2004a, März 20). *PageRank*. https://de.wikipedia.org/wiki/PageRank

[83] Wikipedia-Autoren. (2004b, Juli 2). *Information Retrieval*. https://de.wikipedia.org/wiki/Information_Retrieval

[84] Wikipedia-Autoren. (2007, 24. November). *Compressed Row Storage*. https://de.wikipedia.org/wiki/Compressed_Row_Storage

[85] Wikipedia-Autoren. (2009, 28. Oktober). *Satz von Perron-Frobenius*. https://de.wikipedia.org/wiki/Satz_von_Perron-Frobenius

[86] Xing & Ghorbani. (2004). *Weighted PageRank Algorithm*. Kansas State University. Abgerufen am 18. Dezember 2022, von https://people.cs.ksu.edu/~halmohri/files/weightedPageRank.pdf

12. Selbstständigkeitserklärung

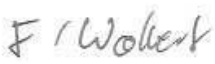
Ich erkläre, dass ich diese besondere Lernleistung mit der Leitfrage

Der PageRank-Algorithmus –

Wie könnte eine geheime Erweiterung des PageRank-Algorithmus von Google aussehen?

Selbstständig angefertigt und keine andren als die von mir angegebenen Quellen und Hilfsmittel verwendet habe. Die den benutzten Werken wörtlich oder inhaltlich entnommenen Stellen sind als solche gekennzeichnet.

Berlin, 19. Dezember 2022



Floyd Wollert