PageRank-Algorithmus

Uhlmann Rudolf

GOOGLE PageRank-Algorithmus

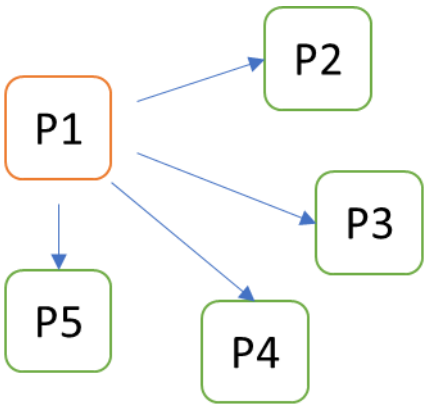
„Der PageRank-Algorithmus ist ein Verfahren, eine Menge verlinkter Dokumente, beispielsweise das World Wide Web, anhand ihrer Struktur zu bewerten und zu gewichten. Dabei wird jedem Element ein Gewicht, den PageRank, aufgrund seiner Verlinkungsstruktur zugeordnet. Der Algorithmus wurde von Larry Page (daher der Name PageRank) und Sergei Brin an der Stanford University entwickelt und von dieser zum Patent angemeldet.[…] Er diente der Suchmaschine Google des von Brin und Page gegründeten Unternehmens Google Inc. als Grundlage für die Bewertung von Seiten.“ (Wikipedia)

Abb. 1 Larry Page und Sergei Brin

Die Grundidee

Webseiten sind über Links miteinander verbunden. Zeigt eine Seite *P1* auf eine Seite *P2*, so vererbt *P1* einen Teil seiner Relevanz (Importance I) an die Seite *P2*.

Beispiel: Seite P_1 mit Relevanz I_1 verweist auf 4 weitere Seiten.



$$I_2 = \frac{I_1}{4}$$

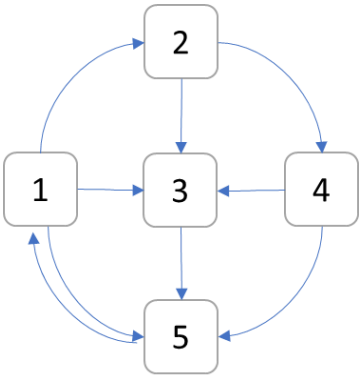
$$I_P = \frac{I_1}{n_1} + \frac{I_2}{n_2} + \cdots$$

Die Relevanz der Seite *P* entsteht durch Übertragung anderer auf sie verweisenden Relevanzen.

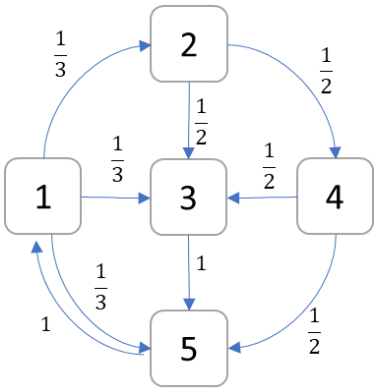
$\frac{1}{n_i}$ … Übergangsfaktoren

Beispiel: MiniWeb (5 Seiten)

Die **LINKSTRUKTUR**



Die sich daraus ergebenden **ÜBERGANGSFAKTOREN**



Daraus ergibt sich das **GLEICHUNGSSYSTEM**

$$
\begin{aligned}
I_1 &= 1 \cdot I_5 \\
I_2 &= \tfrac{1}{3} \cdot I_1 \\
I_3 &= \tfrac{1}{3} \cdot I_1 + \tfrac{1}{2} \cdot I_2 + \tfrac{1}{2} \cdot I_4 \\
I_4 &= \tfrac{1}{2} \cdot I_2 \\
I_5 &= \tfrac{1}{3} \cdot I_1 + 1 \cdot I_3 + \tfrac{1}{2} \cdot I_4
\end{aligned}
\quad \rightarrow \quad
\begin{aligned}
I_1 - I_5 &= 0 \\
-\tfrac{1}{3} I_1 + I_2 &= 0 \\
-\tfrac{1}{3} I_1 - \tfrac{1}{2} I_2 + I_3 - \tfrac{1}{2} I_4 &= 0 \\
-\tfrac{1}{2} I_2 + I_4 &= 0 \\
-\tfrac{1}{3} I_1 - I_3 - \tfrac{1}{2} I_4 + I_5 &= 0
\end{aligned}
$$

Es sind 5 Gleichungen und 5 Variable, trotzdem ist das Gleichungssystem unterbestimmt. Da die Übergangsfaktoren aller ausgehenden Links einer Webseite in Summe 1 ergeben müssen, ist eine der Gleichungen (z.B. die letzte für I_5) überflüssig, da ohnehin aus den anderen ableitbar. Wir können sie also streichen und eine der restlichen Variablen als Vorgabe frei wählen (z.B. $I_1 = 1$).

Die dazu sich ergebenden Lösungen für I_2 bis I_4 und auch für I_5 sind:

$$I_1 = 1 \;\; ; \;\; I_2 = \frac{1}{3} \;\; ; \;\; I_3 = \frac{7}{12} \;\; ; \;\; I_4 = \frac{1}{6} \;\; ; \;\; I_5 = 1$$

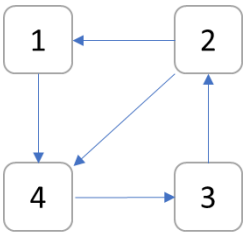
Bemerkungen:

- Das Gleichungssystem ist homogen, die triviale Lösung aller $I_n = 0$ aber uninteressant.
- Gibt es eine nichttriviale Lösung, dann sind auch alle Vielfachen Lösungen.
- Die Lösungen sind von der Wahl der einen Variablen (hier $I_1 = 1$) abhängig, das PageRanking ändert sich dadurch aber nicht!

$$I_1 = I_5 > I_3 > I_2 > I_4$$

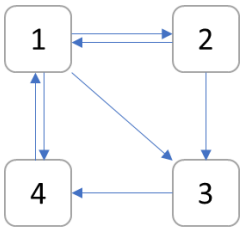
Weitere Beispiele:

a)



Mögliche Lösung: $\left(\frac{1}{7}, \frac{2}{7}, \frac{2}{7}, \frac{2}{7}\right)$

b)

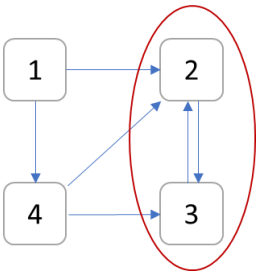


Mögliche Lösung: $\left(\frac{6}{16}, \frac{2}{16}, \frac{3}{16}, \frac{5}{16}\right)$

Der erweiterte PageRank-Algorithmus

„RankSink“-Problem

Beispiel: 2 und 3 spielen nicht mit und vererben keine Relevanzen an die anderen. Dies führt zu nicht brauchbaren Lösungen.



Gleichungssystem:

$$I_1 = \frac{1}{3} I_4$$
$$I_2 = \frac{1}{2} I_1 + I_3 + \frac{1}{3} I_4$$
$$I_3 = I_2 + \frac{1}{3} I_4$$
$$I_4 = \frac{1}{2} I_1$$

I_1	I_2	I_3	I_4	
1	0	0	$-1/3$	0
$-1/2$	1	-1	$-1/3$	0
0	-1	1	$-1/3$	0
$-1/2$	0	0	1	0

Das Gleichungssystem hat abgesehen von der trivialen keine Lösung.

Aus diesem und ähnlich gelagerten Fällen zeigt sich, dass es besser ist, grundsätzlich jeder Seite einen „Basis-PageRank“ zwischen 0 und 1 zuzuweisen.
Der eigentliche PageRank ergibt sich bei Google zu 85% aus dem WWW und zu 15% aus dem Basiswert 1.

Das Gleichungssystem ist jetzt inhomogen!

$$I_1 = 0{,}85 \cdot \left(\frac{1}{3} I_4\right) + 0{,}15 \cdot 1$$
$$I_2 = 0{,}85 \cdot \left(\frac{1}{2} I_1 + I_3 + \frac{1}{3} I_4\right) + 0{,}15 \cdot 1$$
$$I_3 = 0{,}85 \cdot \left(I_2 + \frac{1}{3} I_4\right) + 0{,}15 \cdot 1$$
$$I_4 = 0{,}85 \cdot \left(\frac{1}{2} I_1\right) + 0{,}15 \cdot 1$$

I_1	I_2	I_3	I_4	
1	0	0	$-0{,}85/3$	0,15
$-0{,}85/2$	1	$-0{,}85$	$-0{,}85/3$	0,15
0	$-0{,}85$	1	$-0{,}85/3$	0,15
$-0{,}85/2$	0	0	1	0,15

Lösung: $(0{,}22\ ;\ 1{,}79\ ;\ 1{,}74\ ;\ 0{,}24)$ und somit $I_2 > I_3 > I_4 > I_1$!

Alternative Lösungsvariante

Lösung durch Iteration

Betrachten wir das ursprüngliche Gleichungssystem (ohne Dämpfungsfaktor)

$$I_1 = 1 \cdot I_5$$

$$I_2 = \frac{1}{3} \cdot I_1$$

$$I_3 = \frac{1}{3} \cdot I_1 + \frac{1}{2} \cdot I_2 + \frac{1}{2} \cdot I_4$$

$$I_4 = \frac{1}{2} \cdot I_2$$

$$I_5 = \frac{1}{3} \cdot I_1 + 1 \cdot I_3 + \frac{1}{2} \cdot I_4$$

In der Matrixschreibweise:

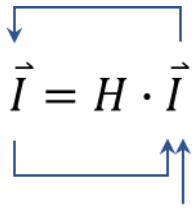
$$\begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 1/3 & 0 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 0 & 1 & 1/2 & 0 \end{pmatrix} \cdot \begin{pmatrix} I_1 \\ I_2 \\ I_3 \\ I_4 \\ I_5 \end{pmatrix}$$

$$\vec{I} = H \cdot \vec{I}$$

H … Linkmatrix

(Mit Dämpfungsfaktor $\vec{I} = (1 - m) \cdot H \cdot \vec{I} + m \cdot \vec{1}$)

Die Idee



Die Gleichung ist rückbezüglich. Die Lösung $\vec{I}$ soll rechts eingesetzt links wieder $\vec{I}$ ergeben. Die Idee beruht jetzt darauf, mit einem beliebigen Startvektor $\vec{I}^{(0)}$ rechts zu beginnen und den daraus erhaltenen Vektor $\vec{I}^{(1)} = H \cdot \vec{I}^{(0)}$ neuerlich rechts einzusetzen. Konvergiert diese Vorgangsweise, so erhält man als Grenzwert die Lösung I : $\vec{I}^{(n)} \xrightarrow{n \to \infty} I$

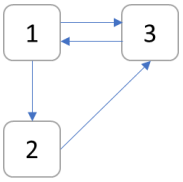
Unser Beispiel:

0)	1)	2)	3)	4)	5)		∞
1	1	1,8	1,9	1,4	1,6		**1,6**
1	0,3	0,3	0,6	0,6	0,5		**0,5**
1	1,3	0,8	0,9	1,0	0,9	…	**0,9**
1	0,5	0,2	0,2	0,3	0,3		**0,3**
1	1,8	1,9	1,4	1,6	1,6		**1,6**

Bemerkungen:

- Konvergiert der Algorithmus immer? NEIN!
- Ist die Lösung unabhängig von $I^{(0)}$? NEIN!
- Allerdings: Der „Verbesserte PageRank-Algorithmus“ konvergiert und die Reihenfolge des PageRankings ist von $I^{(0)}$ unabhängig!

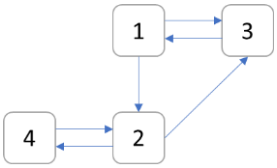
Weiter Beispiele:



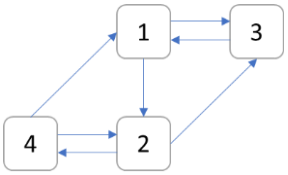
Löse

a) durch lösen des Gleichungssystems
b) durch Iteration

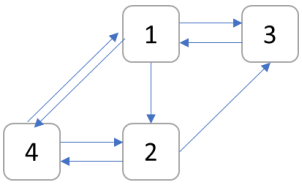
Lsg: (1.15 , 0.66 , 1.19)



Lsg: (1.09 , 1.16 , 1.11 , 0.64)



Lsg: (1.35 , 0.96 , 1.13 , 0.56)



Lsg: (1.3 , 0.9 , 0.9 , 0.9)