

# Inside Google - Algorithmen für Suchmaschinen Page Quality

Alexandros Panagiotidis

Betreuer: Prof. Javier Esparza

Institut für Formale Methoden der Informatik  
Sichere und Zuverlässige Softwaresysteme  
Universität Stuttgart

20. November 2006

## **Zusammenfassung**

Ein kleiner Überblick über das User-Visitation Model[1][2] nach J. Cho und R. E. Adams

## Inhaltsverzeichnis

<b>1</b>	<b>Einleitung</b>	<b>3</b>
<b>2</b>	<b>Qualität einer Seite</b>	<b>4</b>
<b>3</b>	<b>User-Visitation Model</b>	<b>5</b>
3.1	Entwicklung der Popularität . . . . .	5
3.2	Herleitung der Qualitäts-Abschätzung . . . . .	6
3.3	Besser als PageRank? . . . . .	7
3.4	Schöne neue Welt . . . . .	8
<b>4</b>	<b>Fazit</b>	<b>9</b>
	<b>Literatur</b>	<b>10</b>
<b>A</b>	<b>Anhang</b>	<b>11</b>
A.1	Beweis von Gleichung 6 . . . . .	11

## 1 Einleitung

Wer heutzutage nach Informationen sucht, wird früher oder später im Internet nach diesen suchen. Allerdings ist eine solche Suche bei der Geschwindigkeit, mit der das Internet wächst jetzt schon eine Lebensaufgabe.

Aus diesem Grund wurden Suchmaschinen entwickelt, die das Web und andere Teile des Internets durchforsten („crawl“) und für Benutzer indizieren. Jedoch liefern bereits simple Anfragen an eine Suchmaschine so viele Treffer, dass ein Mensch im Allgemeinen unmöglich alle Ergebnisse selbständig auf Relevanz und Qualität analysieren könnte. Daher übernehmen heutige Suchmaschinen auch diesen Teil der Web-Recherche.

Insbesondere Google[3] hat sich in den letzten Jahren als dominierende Suchmaschine in Europa hervorgetan - allein in Deutschland werden über 85% aller Suchanfragen[4] von Google beantwortet.

### Googles Geheimrezept

Dieser Erfolg ist zu grossen Teilen zurückzuführen auf Googles Implementation der Link-Popularitäts-Metrik *PageRank*[5][6]. Der PageRank einer Seite  $u$  ist die gewichtete Summe aller PageRanks der Seiten  $v$ , welche auf  $u$  verlinken, oder formal:

$$PR(u) = (1 - d) + d \sum_{v \in P(u)} \frac{PR(v)}{N_v} \quad (1)$$

mit

- $P(u)$  der Menge aller Seiten, welche auf  $u$  verlinken
- $N_v$  der Anzahl der ausgehenden Links auf Webseite  $v$
- $d$  dem Dämpfungsfaktor bzw. „Langeweile-Faktor“

Eine anschauliche Interpretation des PageRanks bietet das Random Surfer Model[7]: ein Web-Benutzer erreicht eine Seite durch das Folgen von Links auf anderen Seiten<sup>1</sup> oder durch Zufall<sup>2</sup>.

### Rich get richer?

Betrachten wir nun folgendes Szenario. Die Seite  $u_1$  ist seit längerer Zeit online und hat einen hohen PageRank  $PR(u_1)$ . Sie entschliessen sich jetzt eine Seite  $u_2$  zum selben Thema zu erstellen, welche jedoch qualitativ besser sein soll. Da  $u_2$  erst seit Kurzem online ist, existieren noch nicht sehr viele Links darauf.  $PR(u_2)$  ist daher relativ klein und im Speziellen kleiner als  $PR(u_1)$ .

Wenn ein Web-Benutzer nun Seiten zu diesem Thema sucht, so wird Google  $u_1$  vor  $u_2$  in den Suchergebnissen einsortieren. Da Web-Benutzer meist nicht alle Ergebnisse einer Suchanfrage beachten, sind also alle Seiten, welche weit hinten gerankt wurden, benachteiligt. Dieses Phänomen wird mit „rich get richer“ bezeichnet: populäre Seiten werden immer populärer und (momentan) unpopuläre Seiten haben kaum eine Chance populärer zu werden.

---

<sup>1</sup>mit der Wahrscheinlichkeit  $d$

<sup>2</sup>gleichverteilt auf alle Seiten mit der Wahrscheinlichkeit  $1 - d$

## 2 QUALITÄT EINER SEITE

---

Eine experimentelle Untersuchung einer Teilmenge des Webs[2] hat gezeigt, dass dies tatsächlich der Fall zu sein scheint. Es wurden diverse Webseiten mit sieben Monaten Abstand gespiegelt. Anschließend wurde untersucht, wie sich die Popularität verändert hat. Das Ergebnis war, dass populäre Seiten fast alle neuen Links erhalten haben, unpopuläre Seiten nahezu keine neuen Links erhalten haben und manche Seiten sogar Links verloren haben (und dadurch „unpopulärer“ wurden)!

Dieser Makel ist der Grundidee des Verfahrens anzulasten - es wird nur die Struktur der Verlinkung zur Berechnung der Popularität einbezogen. Nur implizit kann der PageRank als Aussage über die *Qualität* einer Seite verwendet werden<sup>3</sup>.

Aber wie will man dieses Problem beheben? Wie misst man die Qualität einer Seite? Ist Qualität nicht etwas sehr Subjektives, was erst durch den Web-Benutzer entschieden werden kann? Im Folgenden wollen wir uns nicht näher mit dem PageRank befassen, sondern widmen uns der Frage, wie die Qualität einer Seite festgestellt werden kann und wie daraus eine neue, bessere Metrik abgeleitet werden kann.

## 2 Qualität einer Seite

Zunächst muss also geklärt werden, was die Qualität einer Seite überhaupt ist. Behalten wir die Idee des PageRanks zunächst bei, ist eine intuitive Definition der Qualität einer Seite  $p$  die Wahrscheinlichkeit, dass ein Web-Benutzer einen Link zu  $p$  erstellt, wenn er die Webseite  $p$  zum ersten Mal wahrnimmt. Mit den Mitteln der Stochastik können wir dies formalisieren:

**Definition** Sei  $A_p$  das Ereignis, dass ein Web-Benutzer  $p$  zum ersten Mal wahrnimmt. Weiter sei  $L_p$  die Wahrscheinlichkeit, dass diesem Web-Benutzer  $p$  so sehr gefällt, dass er einen Link zu  $p$  erstellt. Dann ist die *page quality* der Seite  $p$

$$Q(p) = P(L_p|A_p) \quad (2)$$

Diese Definition ist jedoch nicht praktikabel, da man jede Seite jedem Web-Benutzer zeigen müsste, um  $Q(p)$  für eine Seite  $p$  zu bestimmen. Stattdessen untersuchen wir die Veränderung der Popularität über einen gewissen Zeitraum, um die Qualität einer Seite zu approximieren:

$$Q(p) \approx C \cdot \frac{\Delta PR(p)}{PR(p)} + PR(p) \quad (3)$$

Betrachten wir Gleichung 3 näher.

- $\frac{\Delta PR(p)}{PR(p)}$  beschreibt die relative Änderung der Popularität, gewichtet mit der Konstanten  $C$ .
- Bereits populäre Seiten erhalten i.d.R. wenige neue Links. Daher fließt die bisherige Popularität  $PR(p)$  in die Qualität ein.
- Obwohl hier der PageRank zur Approximation verwendet wird, kann im Grunde jede Popularitäts-Metrik benutzt werden.

---

<sup>3</sup>PageRank kann sogar nur Aussagen über die *Popularität* einer Seite zu einem bestimmten Zeitpunkt machen

Mit dieser Definition wäre es effektiv möglich, die Qualität einer Seite zu messen, z.B. indem man mehrere Snapshots des Webs herunterlädt, die PageRanks berechnet und anschließend  $Q(p)$  bestimmt.

Aber wie kommen wir zu dieser Form von  $Q(p)$  und welche Implikationen stecken hinter dieser Gleichung? Antworten darauf finden wir im *user-visitation model*[1], welches im Folgenden näher beleuchtet wird.

### 3 User-Visitation Model

Das *user-visitation model*[1] von J. Cho und R. E. Adams basiert auf zwei Annahmen:

1. Die Anzahl der Seitenaufrufe einer Seite ist proportional zu ihrer Popularität.
2. Alle Web-Benutzer rufen eine bestimmte Seite mit der selben Wahrscheinlichkeit auf.

Eine Erklärung für Annahme 2 wäre: Wenn einer von  $n$  Web-Benutzern die Seite  $p$  aufruft, so hätte jeder beliebige Web-Benutzer diese Seite mit der Wahrscheinlichkeit  $\frac{1}{n}$  aufrufen können. Um die Plausibilität von Annahme 1 zu erläutern, benötigen wir erst ein paar Begriffe.

**Definition** Die *popularity*  $P(p, t)$  einer Seite  $p$  zum Zeitpunkt  $t$  sei der Anteil der Web-Benutzer, welche  $p$  mögen.

Im Allgemeinen dürfte es so nicht möglich sein  $P(p, t)$  direkt zu bestimmen, daher wird eine andere Metrik (z.B. PageRank) verwendet.

**Definition** Die *visit popularity*  $V(p, t)$  einer Seite  $p$  zum Zeitpunkt  $t$  sei die Anzahl der Aufrufe von  $p$  in einem Zeitraum zum Zeitpunkt  $t$ .

Nach Annahme 1 gilt also

$$V(p, t) = rP(p, t)$$

mit einer Normalisierungskonstanten  $r$ . Dies macht Sinn, da populäre Seiten häufiger als unpopuläre Seiten aufgerufen werden.

#### 3.1 Entwicklung der Popularität

Gehen wir einen Schritt weiter und untersuchen, wie die Popularität einer Seite sich entwickelt. Zunächst halten wir fest, dass die Anzahl der Web-Benutzer, welche eine Seite  $p$  kennen, sich abhängig von der Zeit ändert:

**Definition** Die *user awareness*  $A(p, t)$  einer Seite  $p$  zum Zeitpunkt  $t$  sei der Anteil der Web-Benutzer, welche die Seite  $p$  kennen.

Anders als die Popularität  $P(p, t)$  macht die *user awareness* eine Aussage darüber, wieviele Web-Benutzer  $p$  bereits gesehen haben. Dabei spielt es keine Rolle, ob ihnen  $p$  gefällt oder nicht. Wie man leicht sieht, gilt:

$$P(p, t) = A(p, t)Q(p) \tag{4}$$

Bei der Interpretation dieser Beziehung muss man beachten, dass die Qualität einer Seite, nicht von der Zeit abhängig ist. Wenn sich also die Popularität von  $p$  ändert, dann weil die *user awareness* und nicht die Qualität sich geändert hat!

Weiterhin lässt sich zeigen, dass  $A(p, t)$  sich aus der bisherigen Popularität berechnen lässt:

$$A(p, t) = 1 - e^{-\frac{r}{n} \int_0^t P(p,t) dt} \tag{5}$$

*Beweis.* Die Wahrscheinlichkeit, dass einer von  $n$  Web-Benutzern eine Seite  $p$  nicht aufruft beträgt  $(1 - \frac{1}{n})$ . Wenn  $p$  nun  $k$ -mal aufgerufen wurde, so ist die Wahrscheinlichkeit, dass ein Web-Benutzer  $p$  nicht gesehen hat

$$1 - A(p, t) = \left(1 - \frac{1}{n}\right)^k$$

Nachdem  $t$  Zeit vergangen ist, wurde  $u \int_0^t V(p, t) dt$  mal aufgerufen. Daraus folgt

$$\begin{aligned} 1 - A(p, t) &= \left(1 - \frac{1}{n}\right)^{\int_0^t V(p, t) dt} \\ &= \left(1 - \frac{1}{n}\right)^{r \int_0^t P(p, t) dt} \\ &= \left(\left(1 - \frac{1}{n}\right)^{-n}\right)^{-\frac{r}{n} \int_0^t P(p, t) dt} \end{aligned}$$

Wenn wir von sehr vielen Web-Benutzern ausgehen können wir beobachten

$$\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^{-n} = e$$

und damit

$$1 - A(p, t) = e^{-\frac{r}{n} \int_0^t P(p, t) dt}$$

□

Aus den Gleichungen 4 und 5 folgt schliesslich, dass die Entwicklung der Popularität einer Seite  $p$  berechnet werden kann, wenn man die initiale Popularität  $P(p, 0)$  kennt:

$$P(p, t) = \frac{Q(p)}{1 + \left(\frac{Q(p)}{P(p, 0)} - 1\right) e^{-\frac{r}{n} Q(p) t}} \quad (6)$$

Ein Beweis hierfür ist im Anhang A.1 gegeben. Betrachtet man den Graphen (Abbildung 1) dieser Form von  $P(p, t)$  lässt sich die Popularitätsentwicklung einer Seite in drei Abschnitte einteilen. Zunächst ist die Seite relativ unbekannt und unpopulär. Dann wird die Seite zunehmend bekannter und ihre Popularität steigt. Schliesslich erreicht die Popularität einen Wert, der sich kaum verändert. Insbesondere gilt sogar, dass mit  $t \rightarrow \infty$  die Popularität zu  $Q(p)$  strebt.

### 3.2 Herleitung der Qualitäts-Abschätzung

Erinnern wir uns zurück an Gleichung 3. Um ein berechenbares Verfahren für  $Q(p)$  zu finden, benötigen wir die Veränderung der Popularität. Differenzieren von Gleichung 4 und 5 ergibt

$$Q(p) = \left(\frac{n}{r}\right) \frac{\frac{d}{dt} P(p, t)}{P(p, t) (1 - A(p, t))} \quad (7)$$

Problematisch an dieser Gleichung ist  $A(p, t)$ , da dies nur schwierig zu bestimmen ist und daher nicht weiter beachtet wird. Den Rest bezeichnen wir als *relative popularity increase*

$$I(p, t) = \left(\frac{n}{r}\right) \frac{\frac{d}{dt} P(p, t)}{P(p, t)} \quad (8)$$

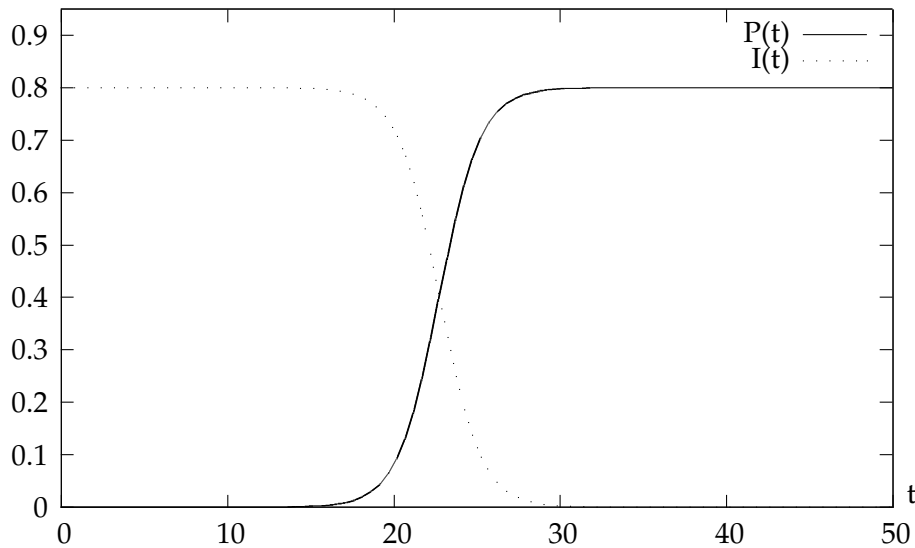


Abbildung 1: Populärkeitsentwicklung und relative Populärkeitsveränderung

Wie man in Abbildung 1 sieht, verhält sich  $I(p, t)$  invers zu  $P(p, t)$ . Die Populärkeit einer Seite kann nur solange steigen, bis schliesslich alle Web-Benutzer eine Meinung über  $p$  haben<sup>4</sup>. Dann ändert sich die Populärkeit nicht mehr. Dieser Beobachtung folgt, dass die momentane Populärkeit  $P(p, t)$  und die Änderung der Populärkeit  $I(p, t)$  zusammen gerade den *quality estimator*  $Q(p)$  ergeben (Abbildung 2), also

$$Q(p, t) = I(p, t) + P(p, t) = \left(\frac{n}{r}\right) \frac{\frac{d}{dt}P(p, t)}{P(p, t)} + P(p, t) \quad (9)$$

Nun können wir effektiv  $Q(p)$  als die Summe der Änderung der Populärkeit und der momentanen Populärkeit berechnen!

### 3.3 Besser als PageRank?

Um zu bestimmen, wie effektiv das user-visitation model als Ranking Methode ist, wurde ein kleines Experiment[1][2] durchgeführt. Zunächst wurden vier Snapshots von diversen Webseiten heruntergeladen und die PageRanks des resultierenden Web-Graphens berechnet. Anschliessend wurde für diejenigen Seiten, deren PageRank nur gestiegen (oder gefallen) ist, die Qualität mit der Formel<sup>5</sup>

$$Q(p) = 0.1 \left( \frac{PR(p, t_3) - PR(p, t_1)}{PR(p, t_1)} \right) + PR(p, t_3)$$

<sup>4</sup>In diesem Model werden Schwankungen der Populärkeit nicht erfasst, daher steigt die Populärkeit lediglich

<sup>5</sup> $C = 0.1$  hat die besten Resultate geliefert

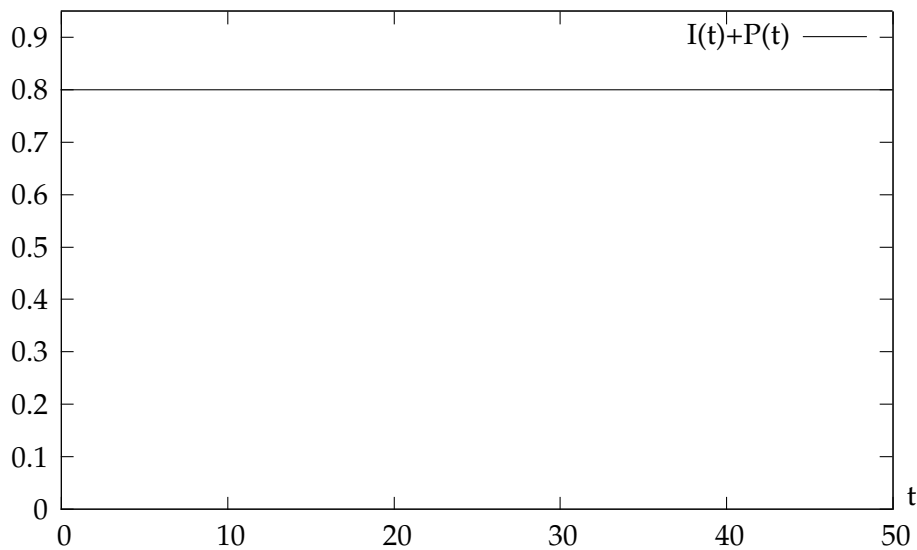


Abbildung 2:  $Q(p) = I(p, t) + P(p, t)$

berechnet. Schliesslich wurde berechnet, wie gut  $Q(p)$  den PageRank zum Zeitpunkt  $t_4$  "vorhergesagt" hat. Dazu wurde folgende Fehlermetrik verwendet:

$$err(p) = \begin{cases} \left| \frac{PR(p, t_4) - Q(p)}{PR(p, t_4)} \right| & \text{für } Q(p) \\ \left| \frac{PR(p, t_4) - PR(p, t_3)}{PR(p, t_4)} \right| & \text{für } PR(p, t_3) \end{cases}$$

Das Ergebnis dieses Experiments war, dass der durchschnittliche Fehler  $err(p) = 0.32$  für  $Q(p)$  und  $err(p) = 0.78$  für  $PR(p, t_3)$  war, d.h.  $Q(p)$  hat den künftigen PageRank  $PR(p, t_4)$  mehr als doppelt so gut bestimmt, wie  $PR(p, t_3)$ !

#### 3.4 Schöne neue Welt

Trotz der offensichtlichen Verbesserung gegenüber dem bisher verwendeten PageRank, gibt es einige Mängel im user-visitation model:

- Im Model nimmt die Popularität von Seiten nur zu. Tatsächlich kann die Popularität von Seiten aber auch abnehmen. Dazu müsste man modellieren, dass bereits besuchte Seiten auch "vergessen" werden können.
- Bei Seiten, deren PageRank osziliert, ist es schwierig  $I(p, t)$  zu bestimmen. Der Einfachheit halber wird für diese Seiten  $I(p, t) = 0$  gesetzt.
- Für neue Seiten sollte die Änderung der Popularität über einen längeren Zeitraum gemessen werden, als für Seiten, welche länger existieren.
- Das Experiment zur Validierung des Models bestand nur aus einem relativ kleinem Teil des Webs. Ob die Ergebnisse ähnlich gut sind, wenn man eine größere Menge von Seiten betrachtet, ist noch unklar.

Als weiterführende Arbeit wäre es möglich, statt der Änderung des PageRanks die Änderung der Aufrufe zu beobachten und zu vergleichen, welche Methode bessere Ergebnisse liefert.

## 4 Fazit

Wir haben gesehen, dass der PageRank bereits ein gutes Verfahren zur Anordnung von Suchergebnissen ist. Allerdings betrachtet PageRank nur die aktuelle Struktur des Webs. Daher haben neue Seiten mit hoher Qualität kaum eine Chance, populär zu werden.

Das user-visitation model versucht, dieses Problem in den Griff zu bekommen, indem nicht die Link-Struktur zu einem Zeitpunkt betrachtet wird. Stattdessen wird die Veränderung der Verlinkungen verwendet, um eine Seite zu ranken.

Experimentell hat dieses Model den künftigen PageRank einer Seite doppelt so gut bestimmt, wie PageRank selbst.

Es bleibt abzuwarten, ob diese Methode sich durchsetzen wird als Ranking Algorithmus der dritten Generation.

### Literatur

- [1] J. Cho and R. E. Adams. Page quality: In search of an unbiased web ranking. Technical report, UCLA Computer Science Department, November 2003.
- [2] J. Cho and S. Roy. Impact of search engines on page popularity. Technical report, UCLA Computer Science Department, May 2004.
- [3] *Google*, <http://www.google.com/>
- [4] *WebHits*, <http://www.webhits.de/deutsch/index.shtml?webstats.html>
- [5] *PageRank Artikel bei Wikipedia*, <http://de.wikipedia.org/wiki/PageRank>
- [6] *Google Technology*, <http://www.google.com/technology/>
- [7] *GetStyle.net Google PageRank*, <http://www.getstyle.net/pagerank/>
- [8] *Logistische Gleichung bei Wikipedia*, [http://de.wikipedia.org/wiki/Logistische\\_Gleichung](http://de.wikipedia.org/wiki/Logistische_Gleichung)

## A Anhang

### A.1 Beweis von Gleichung 6

*Beweis.* Nach Gleichung 4 und 5 gilt

$$P(p, t) = A(p, t)Q(p) = \left(1 - e^{-\frac{r}{n} \int_0^t P(p, t) dt}\right) Q(p)$$

Mit  $f(t) = e^{-\frac{r}{n} \int_0^t P(p, t) dt}$  gilt

$$\begin{aligned} \ln f &= -\frac{r}{n} \int_0^t P(p, t) dt \\ \left(\frac{1}{f}\right) \frac{df}{dt} &= \left(-\frac{r}{n}\right) P(p, t) \\ \left(-\frac{n}{r}\right) \left(\frac{1}{f}\right) \frac{df}{dt} &= P(p, t) \end{aligned}$$

und somit

$$\left(-\frac{n}{r}\right) \left(\frac{1}{f}\right) \frac{df}{dt} = (1 - f) Q(p) \quad (10)$$

Gleichung 10 ist eine sogenannte Verhulst-Gleichung (auch logistische Gleichung[8]). Die Lösung dieser Gleichung ist

$$f = \frac{1}{1 + C e^{\frac{r}{n} Q(p) t}}$$

Im Folgenden wollen wir die Konstante C bestimmen:

$$\begin{aligned} e^{-\frac{r}{n} \int_0^t P(p, t) dt} &= \frac{1}{1 + C e^{\frac{r}{n} Q(p) t}} \\ -\frac{r}{n} \int_0^t P(p, t) dt &= -\ln \left(1 + C e^{\frac{r}{n} Q(p) t}\right) \\ \left(-\frac{r}{n}\right) P(p, t) &= -\frac{1}{1 + C e^{\frac{r}{n} Q(p) t}} \left(\frac{r}{n}\right) Q(p) C e^{\frac{r}{n} Q(p) t} \\ P(p, t) &= \frac{Q(p) C e^{\frac{r}{n} Q(p) t}}{1 + C e^{\frac{r}{n} Q(p) t}} \\ &= \frac{Q(p) C}{e^{-\frac{r}{n} Q(p) t} + C} \end{aligned} \quad (11)$$

Mit  $t = 0$  in Gleichung 11 können wir nach C auflösen:

$$\begin{aligned} P(p, 0) &= \frac{Q(p) C}{C + 1} \\ CP(p, 0) + P(p, 0) &= Q(p) C \\ C &= \frac{P(p, 0)}{Q(p) - P(p, 0)} \end{aligned}$$

Schliesslich ergibt sich

$$\begin{aligned} P(p, t) &= \frac{\frac{P(p,0)}{Q(p)-P(p,0)}Q(p)}{e^{-\frac{r}{n}Q(p)t} + \frac{P(p,0)}{Q(p)-P(p,0)}} \\ &= \frac{P(p,0)Q(p)}{(Q(p) - P(p,0)) \left( e^{-\frac{r}{n}Q(p)t} + \frac{P(p,0)}{Q(p)-P(p,0)} \right)} \\ &= \frac{Q(p)}{1 + \left( \frac{Q(p)}{P(p,0)} - 1 \right) e^{-\frac{r}{n}Q(p)t}} \end{aligned}$$

□