Eine Anwendung der Theorie der Modulfunktionen in der Informatik

Von

Peter Kirschenhofer und **Helmut Prodinger**

(Vorgelegt in der Sitzung der math.-nat. Klasse am 13. Oktober 1988 durch das w.M. Edmund Hlawka)

1. Einleitung

Die Analyse von Algorithmen hat sich in den letzten Jahren zu einem bedeutenden Teilgebiet der Theoretischen Informatik entwickelt. Neben der Untersuchung von Laufzeit und Speicherbedarf eines Algorithmus im ungünstigsten Fall („Worst-case-Verhalten") erweist sich für die Praxis die Feststellung des „durchschnittlichen" Verhaltens als mehr und mehr bedeutsam. Vom Standpunkt des Mathematikers aus ist dieses Aufgabengebiet besonders reizvoll, da die (asymptotische) Analyse des durchschnittlichen Verhaltens zumeist die Verbindung mathematischer Methoden aus unterschiedlichen Teilgebieten, wie Kombinatorik, Wahrscheinlichkeitstheorie, Funktionentheorie u. a., erfordert. In der gegenständlichen Arbeit soll eine Aufgabenstellung behandelt werden, zu deren Lösung auf Ramanujan zurückgehende Transformationsresultate aus der Theorie der Modulfunktionen von zentraler Bedeutung sind. Wir geben eine kurze Übersicht über eine Reihe von Resultaten, die in letzter Zeit von uns mit dieser Idee erzielt werden konnten, und beweisen ein technisch sehr aufwendiges neues Resultat, welches auf eine Aufgabenstellung von Knuth [11] zurückgeht.

2. Modulfunktionen und Reihentransformationen

Zur Motivation des Einsatzes von Resultaten aus der Theorie der Modulfunktionen sei die folgende Aufgabe erwähnt (die zusammen mit Schoißengeier in [8] behandelt wurde):

Bei der asymptotischen Analyse der Varianz eines Parameters einer speziellen Datenstruktur (vgl. Abschnitt 3) tritt folgende Konstante auf:

$$C_M = \frac{1}{12} + \frac{\pi^2}{6\log^2 M} - \frac{2}{\log M} \cdot \mu_M \tag{2.1}$$

mit

$$\mu_M = \sum_{k \geqq 1} \frac{(-1)^{k-1}}{k\,(M^k - 1)} \qquad (M \geqq 2,\ m \in \mathbb{N}).$$

Die numerische Auswertung ergibt

$$|C_2 - 1| < 10^{-12}. \tag{2.2}$$

Auf der Suche nach einer möglichen Erklärung für diese numerische Koinzidenz und nach einer expliziten Formel für den Fehler gelangt man zur folgenden bei Ramanujan [14] auftretenden Transformationsformel:

Seien $\alpha, \beta > 0$ mit $\alpha \cdot \beta = \pi^2$. Dann gilt

$$\begin{aligned} &\sum_{k=1}^{\infty} \frac{1}{k\,(e^{2\alpha k} - 1)} - \frac{1}{4}\log \alpha + \frac{\alpha}{12} = \\ &= \sum_{k=1}^{\infty} \frac{1}{k\,(e^{2\beta k} - 1)} - \frac{1}{4}\log \beta + \frac{\beta}{12}. \end{aligned} \tag{2.3}$$

Tatsächlich ist nun (2.3) äquivalent zu einem Spezialfall der *Dedekindschen* Transformationsformel für die η-Funktion

$$\eta(\tau) = e^{\pi i \tau/12} \cdot \prod_{n=1}^{\infty} (1 - e^{2\pi i n \tau}), \quad \tau \in H, \tag{2.4}$$

(H die obere komplexe Halbebene), nämlich

$$\eta\left(-\frac{1}{\tau}\right) = (-i\tau)^{1/2} \cdot \eta(\tau). \tag{2.5}$$

Für die Formeln (2.3) bzw. (2.5) existiert ein sehr eleganter Beweis von C. L. Siegel ([16]) mit Hilfe des Residuenkalküls; vgl. auch [1].

Die Lösung des oben erwähnten Problems ist mit (2.3) recht einfach:

Setzen wir

$$f(x) = \sum_{k \geqq 1} \frac{(-1)^{k-1}}{k\,(e^{kx}-1)} = \log \prod_{n \geqq 1} (1 + e^{-nx}), \tag{2.6}$$

so ist

$$\mu_M = f(\log M). \tag{2.7}$$

Wegen

$$\prod_{n \geqq 1} (1 + q^n) = \prod_{n \geqq 0} \frac{1}{1 - q^{2n+1}} \tag{2.8}$$

ist nun

$$f(x) = g(x) - g(2x) \tag{2.9}$$

mit

$$g(x) = -\log \prod_{n \geqq 1} (1 - e^{-nx}) = \sum_{k \geqq 1} \frac{1}{k\,(e^{kx}-1)}. \tag{2.10}$$

Mittels (2.3) folgt dann

$$f(x) = \frac{\pi^2}{12x} - \frac{\log 2}{2} + \frac{x}{24} - f\left(\frac{2\pi^2}{x}\right) \tag{2.11}$$

und damit

$$C_M = \log_M 2 - \frac{2}{\log M} f\left(\frac{2\pi^2}{\log M}\right). \tag{2.12}$$

Mit (2.12) kann nun $C_2 - 1$ leicht abgeschätzt werden.

Dient die Anwendung der Transformationsformel (2.3) in diesem Beispiel nur der Erklärung einer numerischen Koinzidenz, so haben wir es im weiteren mit Aufgabenstellungen zu tun, in denen Formel (2.3) bzw. ähnliche Resultate bei der Feststellung der korrekten asymptotischen Größenordnung vorgegebener Ausdrücke entscheidend eingehen. Die Transformationsformeln, auf die wir uns beziehen werden, behandeln Reihen der Form

$$\sum_{k \geqq 1} \frac{k^m}{e^{2\alpha k} - 1}, \quad m \text{ eine ungerade ganze Zahl, } \alpha > 0: \tag{2.13}$$

Seien $\alpha, \beta > 0$ und $\alpha \cdot \beta = \pi^2$. Dann gilt:
(I) [Ramanujans Formel für $\zeta(2N+1)$]:
Ist N eine positive ganze Zahl, so ist

$$\alpha^{-N} \cdot \left\{\frac{1}{2}\zeta(2N+1) + \sum_{k \geqq 1} \frac{k^{-2N-1}}{e^{2\alpha k}-1}\right\} =$$
$$= (-\beta)^{-N} \cdot \left\{\frac{1}{2}\zeta(2N+1) + \sum_{k \geqq 1} \frac{k^{-2N-1}}{e^{2\beta k}-1}\right\} - \tag{2.14}$$
$$- 2^{2N} \cdot \sum_{k=0}^{N+1} (-1)^k \cdot \frac{B_{2k}}{(2k)!} \cdot \frac{B_{2N+2-2k}}{(2N+2-2k)!} \alpha^{N+1-k} \beta^k,$$

wobei B_n die durch

$$\sum_{n \geqq 0} \frac{B_n z^n}{n!} = \frac{z}{e^z - 1}$$

definierten *Bernoulli*zahlen sind

(II) Für $m = -1$ haben wir bereits Formel (2.3).

(III) Für $m = 1$ gilt:

$$\alpha \cdot \sum_{k \geqq 1} \frac{k}{e^{2\alpha k}-1} + \beta \cdot \sum_{k \geqq 1} \frac{k}{e^{2\beta k}-1} = \frac{\alpha+\beta}{24} - \frac{1}{4}. \tag{2.15}$$

(IV) Für natürliche Zahlen $N \geqq 2$ ist

$$\alpha^N \cdot \sum_{k \geqq 1} \frac{k^{2N-1}}{e^{2\alpha k}-1} + \beta^N \cdot \sum_{k \geqq 1} \frac{k^{2N-1}}{e^{2\beta k}-1} = (\alpha^N - (-\beta)^N) \frac{B_{2N}}{4N}. \tag{2.16}$$

Beweise von (I)—(IV) mit Hilfe der Theorie der modularen Funktionen sowie ausführliche historische Hinweise findet man in der ausgezeichneten Arbeit [2] von B. Berndt.

In den von uns betrachteten Problemstellungen besteht ein zentraler Teil der Lösung nun in der Untersuchung des Mittels des Quadrates von periodischen Funktionen wie etwa

$$\delta_1(x) = \frac{1}{L} \cdot \sum_{k \neq 0} \chi_k \cdot \Gamma(-1-\chi_k) e^{2k\pi i x}, \tag{2.17}$$

mit $L = \log 2$ und $\chi_k = \dfrac{2k\pi i}{L}$.

Der nullte Fourierkoeffizient $[\delta_1^2]_0$ von $\delta_1^2(x)$ kann nun unter Benützung von (2.14) und (2.3) bestimmt werden (vgl. [5]):

Es ist wegen der Funktionalgleichung für $\Gamma(x)$ und wegen

$$|\Gamma(iy)|^2 = \frac{\pi}{y \cdot \sinh(\pi y)},$$

$$\sum_{k \neq 0} |\chi_k|^2 |\Gamma(-1 + \chi_k)|^2 =$$

$$= L \cdot \sum_{j \geqq 0} (-1)^j \left(\frac{L^2}{4\pi^2}\right)^j \cdot 2(g_j(\alpha) - g_j(2\alpha)),$$

wobei

$$\alpha = \frac{\pi^2}{L} \text{ und } g_j(\alpha) = \sum_{l \geqq 1} \frac{1}{l^{2j+1}(e^{2\alpha l} - 1)}.$$

Vermöge der Transformationsformeln erhält man einen wesentlich komplizierteren Ausdruck für die rechts stehende Reihe, der insbesondere Faltungssummen über die Bernoullizahlen enthält. Mit Hilfe einiger trickreichen Manipulationen mit erzeugenden Funktionen vereinfacht sich jedoch das Ergebnis wesentlich, und man erhält letztlich

$$[\delta_1^2]_0 = 3 - \frac{1}{L} - \frac{1}{L^2} + \frac{2}{L} \sum_{k \geqq 2} \frac{(-1)^k k}{(k+1)(k-1)(2^k - 1)}. \tag{2.18}$$

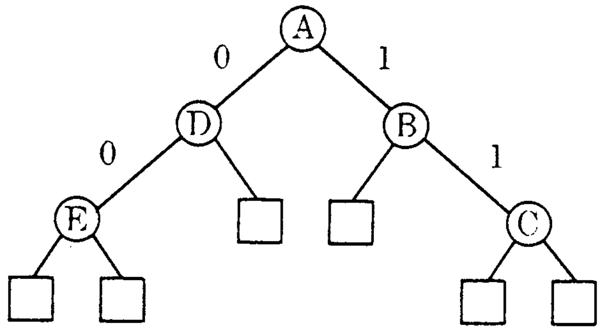
3. Analyse von Datenstrukturen zur digitalen Suche

Eine wichtige Klasse von Algorithmen befaßt sich mit dem Abspeichern und Aufsuchen von Daten in baumartigen Datenstrukturen (vgl. [11]), wobei wir in dieser Arbeit sogenannte *Digitale Suchbäume* untersuchen:

Vorgegeben seien N Schlüssel $K_1, \ldots, K_N$ bzw. deren interne Darstellung durch eine (potentiell unendliche) Binärfolge. Der zugehörige Digitale Suchbaum ist ein Binärbaum, dessen N interne Knoten die Schlüssel $K_1, \ldots, K_N$ beinhalten. Der Baum ist folgendermaßen aufgebaut: K_1 ist in der Wurzel abgespeichert; wurden $K_1, \ldots, K_i$ bereits abgespeichert, dann findet man den Platz für $K_{i+1} = a_1 a_2 a_3 \ldots (a_j \in \{0, 1\})$ so, daß man die Symbole a_j jeweils als

Anweisung betrachtet, in der j-ten Ebene nach links (0) bzw. rechts (1) weiterzugehen, bis zum erstenmal ein noch unbesetzter Knoten auftritt.

Beispiel: A = 010...
B = 110...
C = 111...
D = 001...
E = 000...

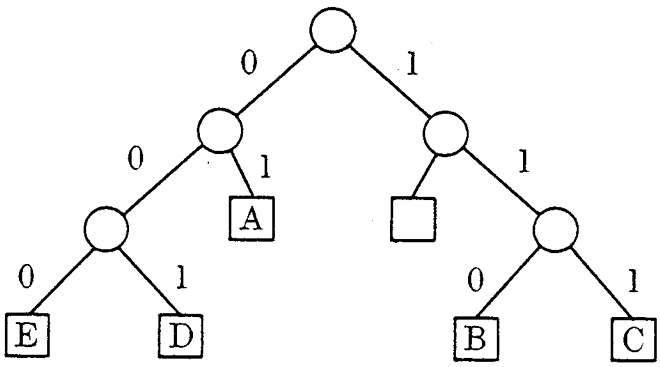


Die Reihenfolge, in der die Schlüssel eingetragen werden, ist natürlich relevant.

Wir werden im weiteren auch Resultate über andere Datenstrukturen zur digitalen Suche zitieren, nämlich sogenannte ,,Tries" (von information re*trie*val) und ,,Patricia Tries" (von *p*ractical *a*lgorithm *t*o *r*etrieve *i*nformation *c*oded *i*n *a*lphanumeric).

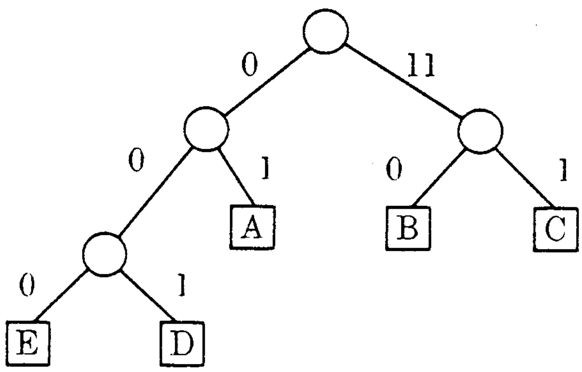
Binäre Tries entstehen in analoger Weise zu digitalen Suchbäumen, jedoch werden die Daten in den Endknoten abgespeichert.

Beispiel:



Patricia Tries entstehen im wesentlichen durch Weglassen derjenigen internen Knoten in Tries, die nur einen Nachfolger haben.

Beispiel:



In den folgenden Ergebnissen wird stets von einer der obigen Datenstrukturen ausgegangen, die aus N Schlüsseln entstanden ist, wobei wir annehmen, daß alle Binärfolgen als Schlüssel gleich wahrscheinlich sind. (Man kann anstelle von Binärfolgen auch Folgen über einem endlichen Alphabet der Größe M betrachten und gelangt dann zu den analogen M-ären Datenstrukturen; vgl. [6].)

In der Informatik interessant ist nun das ,,durchschnittliche Verhalten" gewisser Parameter, die für den Speicherbedarf der Datenstruktur, für den Aufwand bei der Suche bzw. bei der Einfügung von neuen Daten oder ähnlichen Größen charakteristisch sind. Die Beschreibung erfolgt meist durch den *Erwartungswert* des entsprechenden Parameters unter den aus N Schlüsseln gebildeten Bäumen. Zahlreiche Resultate in dieser Richtung finden sich bei Knuth [11] sowie, unter Verwendung neuartiger analytischer Methoden, bei Flajolet und Sedgewick [3]. Für die Praxis ist natürlich neben der Kenntnis des Erwartungswertes eine genauere Information über die Verteilung der zufälligen Veränderlichen wünschenswert: man wird also versuchen, die *Varianz* zu analysieren. Um die dabei auftretenden mathematischen Probleme besser zu verstehen, wollen wir zunächst einen kurzen Abriß der *zur asymptotischen Analyse verwendeten Methode* geben:

Es bezeichne X_N die zu untersuchende zufällige Veränderliche, z. B. die Anzahl der internen Knoten eines aus N Schlüsseln aufgebauten Tries. Im ersten Schritt stellt man eine Rekursionsformel für die zu den X_N gehörigen wahrscheinlichkeitserzeugenden Funktionen auf; diese Rekursionsformel spiegelt im wesentlichen die rekursive Struktur der auftretenden Bäume wider. Im nächsten Schritt gewinnt man eine Rekursion für diese Erwartungswerte EX_N, deren explizite Lösung stets auf die Gestalt

$$EX_N = \sum_{k \geqq m} \binom{N}{k} (-1)^k f(k) \tag{3.1}$$

gebracht werden kann. Man sucht nun nach einer analytischen Fortsetzung der Funktion f, die es erlaubt, das folgende Lemma anzuwenden (das Auffinden einer geeigneten analytischen Fortsetzung wird sich in dem von uns in Abschnitt 4 und 5 behandelten Problem als durchaus nichttriviale Aufgabe erweisen).

Lemma 1 (vgl. [13]): Sei $\mathscr{C}$ eine Kurve, die die Punkte $m, m+1, \ldots, N$ umfährt, und sei $f(z)$ analytisch im Innern von $\mathscr{C}$. Dann gilt

$$\sum_{k \geqq m} \binom{N}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_{\mathscr{C}} [N; z] f(z)\, dz,$$

wobei

$$[N;z] = \frac{(-1)^{N-1} \cdot N!}{z(z-1)\dots(z-N)}. \qquad \square$$

Zumeist ist f nun eine in $\mathbb{C}$ meromorphe Funktion, deren Wachstumseigenschaften im Unendlichen es erlauben, durch Ausdehnung des Integrationsweges den folgenden asymptotischen Ausdruck zu erhalten:

$$EX_N \sim \sum \operatorname{Res}([N;z] f(z)), \tag{3.2}$$

wobei die Summe über alle Pole $\neq m, \dots, N$ erstreckt wird (vgl. [3] und [15] für die technischen Details der geeigneten Ausdehnung des Integrationspfades).

In den von uns oben angesprochenen Aufgaben ergibt sich für den Erwartungswert ein asymptotischer Ausdruck der Form

$$EX_N \sim N \cdot (A + \delta(\log_2 N)) \tag{3.3}$$

bzw.

$$EX_N \sim N \log N \cdot (A + \delta(\log_2 N)),$$

mit einer Konstanten A und einer stetigen, periodischen Funktion $\delta(x)$ mit Mittel 0, deren Bauform ähnlich wie (2.17) ist.

Versucht man nun die Varianz asymptotisch auszuwerten, so verwendet man

$$\operatorname{Var} X_N = EX_N(X_N - 1) + EX_N - (EX_N)^2 \tag{3.4}$$

und geht zur Bestimmung des Moments 2. Ordnung ähnlich wie beim Erwartungswert vor, wobei allerdings weit höhere technische Schwierigkeiten auftreten. In (3.4) geht jedoch mit dem Term $(EX_N)^2$ auch das *Quadrat* $\delta^2(\log_2 N)$ *der periodischen Fluktuation* aus (3.3) ein: Es erweist sich nun als entscheidender Schritt, das Mittel der Funktion $\delta^2(x)$ zu bestimmen (vgl. Abschnitt 2) — hier werden die *Transformationsformeln* aus der Theorie der modularen Funktionen verwendet. Hat man das Mittel als eigenen Term isoliert, so verschwinden alle nichtfluktuierenden Bestandteile der Ordnung $N^2 \cdot \log^2 N$ in $\operatorname{Var} X_N$; das Verschwinden der fluktuierenden Bestandteile ergibt sich dann aus einem Stetigkeitsargument und der Nichtnegativität von $\operatorname{Var} X_N$.

Insgesamt kann man also unter Anwendung der Transformationsformeln zeigen, daß $\operatorname{Var} X_N$ asymptotisch deutlich kleiner ist als vielleicht erwartet. Mit hohem technischen Aufwand können schließlich die tatsächlichen asymptotischen Hauptterme von $\operatorname{Var} X_N$ ermittelt werden, wobei in die „expliziten" Formeln meist sehr komplizierte Konstanten eingehen.

Die Varianzen $\operatorname{Var} X_N$ für die folgenden Parameter bzw. Datenstrukturen wurden bereits studiert: Die *Anzahl der internen Knoten* in Tries ([5]), die *Kosten der Einfügung* eines neuen Datums für alle drei erwähnten Datenstrukturen ([7]) und die sogenannte *Pfadlänge*, das ist die Summe der Abstände der belegten Knoten von der Wurzel, in Tries ([1]), Patricia Tries ([9]) und Digitalen Suchbäumen (in Vorbereitung). Folgen die Arbeiten wohl ein und derselben oben dargelegten Leitlinie, so sind die zu bewältigenden technischen Probleme meist völlig unterschiedlich geartet. In der vorliegenden Arbeit wollen wir nun eine technisch besonders aufwendige Aufgabe bearbeiten, nämlich die Untersuchung der Anzahl der *internen Endknoten* in Digitalen Suchbäumen (d. h. der Endknoten des Baumes, der aus den besetzten Knoten gebildet wird). Knuth hat in [11] die asymptotische Auswertung des Erwartungswertes als offenes Problem gestellt; Flajolet und Sedgewick konnten das Problem in [3] lösen. Wir präsentieren in den folgenden Abschnitten die Analyse der Varianz. Charakteristisch für die Analyse von Digitalen Suchbäumen ist das Auftreten von Funktionen in Form von Produkten, wie etwa

$$Q(x) = \prod_{i \geqq 1} \left(1 - \frac{x}{2^i}\right) \tag{3.5}$$

im einfachsten Fall und die vielfältige Verwendung der auf Euler zurückgehenden Identitäten (4.15).

Wir verwenden im weiteren häufig die folgenden Abkürzungen:

$$L = \log 2, \ \chi_k = \frac{2 k \pi i}{L} \tag{3.6}$$

sowie

$$[z^n] f(z)$$

für den Koeffizienten von z^n in der Laurent-Entwicklung von $f(z)$.

4. Erwartungswert der Anzahl der internen Endknoten

Zur Berechnung der Erwartungswerte und Varianzen erweist es sich als zweckmäßig, die folgenden wahrscheinlichkeitserzeugenden Funktionen $F_N(z)$ zu betrachten:

Der Koeffizient von z^k in $F_N(z)$ sei die Wahrscheinlichkeit dafür, daß ein digitaler Suchbaum aus N Daten genau k interne Endknoten besitzt. Wir erhalten dann die folgende Rekursion

$$F_{N+1}(z) = \sum_{k=0}^{N} 2^{-N} \binom{N}{k} F_k(z) F_{N-k}(z), \qquad N \geqq 1,\ F_0(z) = 1,\ F_1(z) = z. \tag{4.1}$$

[Man beachte, daß $2^{-N}\binom{N}{k}$ die Wahrscheinlichkeit dafür angibt, daß ein digitaler Suchbaum aus $N+1$ Daten k (resp. $N-k$) dieser Daten im linken (resp. rechten) Subbaum enthält.]

Der Erwartungswert l_N des betrachteten Parameters ergibt sich dann als

$$l_N = F_N'(1); \tag{4.2}$$

aus (4.1) erhält man unmittelbar

$$l_{N+1} = 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k,\ N \geqq 1,\ l_0 = 0,\ l_1 = 1. \tag{4.3}$$

Um eine explizite Lösung dieser Rekursion zu gewinnen, erweist sich die Betrachtung der exponentiellen erzeugenden Funktion

$$L(z) = \sum_{N \geqq 0} l_N \frac{z^N}{N!} \tag{4.4}$$

als zweckmäßig. Die Rekursion (4.3) übersetzt sich dann in die Funktionaldifferentialgleichung

$$L'(z) = 1 + 2L\left(\frac{z}{2}\right) e^{z/2}. \tag{4.5}$$

Betrachten wir anstelle von $L(z)$

$$\hat{L}(z) := e^{-z} L(z) = \sum_{N \geqq 0} \hat{l}_N \frac{z^N}{N!}, \tag{4.6}$$

ergibt sich die Beziehung

$$\hat{L}'(z) + \hat{L}(z) = e^{-z} + 2\hat{L}\left(\frac{z}{2}\right)$$

und somit

$$\hat{l}_{N+1} = (-1)^N - (1 - 2^{1-N})\,\hat{l}_N,\; N \geqq 0,\; \hat{l}_0 = 0.$$

Die letzte Rekursion wird durch Iteration gelöst: Bezeichnet Q_N das Produkt

$$Q_N = \prod_{1 \leqq k \leqq N} \left(1 - \frac{1}{2^k}\right), \tag{4.7}$$

so ist

$$\hat{l}_N = (-1)^{N-1} Q_{N-2} \left(\frac{1}{Q_0} + \ldots + \frac{1}{Q_{N-2}}\right),\; N \geqq 2,\; \hat{l}_0 = 0,\; \hat{l}_1 = 1. \tag{4.8}$$

Zur Abkürzung setzen wir

$$R_N = Q_N \left(\frac{1}{Q_0} + \ldots + \frac{1}{Q_N}\right) \tag{4.9}$$

und haben schließlich

$$l_N = N - \sum_{k \geqq 2} \binom{N}{k} (-1)^k R_{k-2}. \tag{4.10}$$

Um Lemma 1 anwenden zu können, ist eine geeignete analytische Fortsetzung der Größen R_N vonnöten. Flajolet und Sedgewick verwenden in [3]

$$R_N = N + 1 - \alpha + R_N^*,\; \alpha = \sum_{k \geqq 1} \frac{1}{2^k - 1}, \tag{4.11}$$

wobei R_N^* fortgesetzt wird durch die Funktion

$$R^*(z) = \sum_{j \geqq 0} \frac{(z + 1 + j - \alpha) \cdot 2^{-z-1-j}}{(1 - 2^{-z-1})(1 - 2^{-z-2}) \ldots (1 - 2^{-z-1-j})}. \tag{4.12}$$

Für unsere weiteren Betrachtungen erweist es sich als zweckmäßiger, die folgende alternative Darstellung zu verwenden:

$$R^*(z) = \frac{1}{Q_\infty} \cdot \sum_{j \geqq 1} \frac{(-1)^{j-1}}{2^{\binom{j}{2}} Q_{j-1}} \cdot \frac{z + j}{2^{z+j} - 1}. \tag{4.13}$$

Hierbei bezeichnet Q_∞ das unendliche Produkt

$$Q_\infty = \prod_{j \geqq 1} \left(1 - \frac{1}{2^j}\right). \tag{4.14}$$

Zum Beweis schreiben wir $R^*(z)$ nach (4.12) mit $x = 2^{-z}$ in der Form

$$R^*(z) = \sum_{j \geqq 1} \frac{(z + j - \alpha) \cdot 2^{-z-j}}{\left(1 - \frac{x}{2}\right)\left(1 - \frac{x}{4}\right) \dots \left(1 - \frac{x}{2^j}\right)}$$

und erhalten durch Partialbruchzerlegung und Vertauschung der Summen

$$R^*(z) = \sum_{j \geqq 1} \frac{(-1)^{j-1}}{2^{\binom{j}{2}} Q_{j-1} \cdot \left(1 - \frac{x}{2^j}\right)} \sum_{l \geqq 0} \frac{z + l + j - \alpha}{Q_l} \cdot 2^{-z-j-l}.$$

Zur Auswertung der inneren Reihe bedient man sich der bekannten Identität von Euler:

$$\sum_{n \geqq 0} \frac{t^n}{(1-q)(1-q^2)\dots(1-q^n)} = \frac{1}{(1-t)(1-qt)(1-q^2t)\dots}. \tag{4.15}$$

Setzt man $q = t = 1/2$, erhält man

$$\sum_{l \geqq 0} \frac{2^{-l}}{Q_l} = \frac{1}{Q_\infty} \tag{4.16}$$

sowie (nach vorherigem Differenzieren)

$$\sum_{l \geqq 0} \frac{l \cdot 2^{-l}}{Q_l} = \frac{\alpha}{Q_\infty}. \tag{4.17}$$

Damit ergibt sich der Ausdruck (4.13) für $R^*(z)$ unmittelbar.

Aus (4.10), (4.11) und (4.13) ergibt sich insgesamt

$$l_N = (N-1)(\alpha+1) - \sum_{k \geqq 2} \binom{N}{k} (-1)^k f_1(k) \tag{4.18}$$

mit

$$f_1(z) = R^*(z-2) = \frac{1}{Q_\infty} \cdot \sum_{j \geqq 1} \frac{(-1)^{j-1}}{2^{\binom{j}{2}} Q_{j-1}} \cdot \frac{z+j-2}{2^{z+j-2}-1}. \tag{4.19}$$

Zur asymptotischen Auswertung von l_N in der späterhin nötigen Genauigkeit sind die Residuen von $[N; z] f_1(z)$ in den (einfachen!) Polen $z = 1 + \chi_k$ bzw. $z = \chi_k$ zu berechnen.

Diese Residuenberechnungen sind nicht schwierig und führen direkt zum Satz 1. Dazu ist noch anzumerken, daß $R^*(-1)$ existiert, was aus der Darstellung (4.13) sofort, aus der Darstellung (4.12) hingegen nur mit Mühe zu erhalten ist. Weiters ist in [3] die Rekursion

$$R^*(z) = \frac{(z+1-\alpha)\, 2^{-z-1}}{1-2^{-z-1}} + \frac{1}{1-2^{-z-1}} R^*(z+1) \tag{4.20}$$

gezeigt worden, woraus man

$$R^*(-2) = 2(\alpha+1) - R^*(-1) \tag{4.21}$$

ersieht.

Satz 1 [Flajolet, Sedgewick]: Der Erwartungswert l_N der Anzahl der internen Endknoten eines digitalen Suchbaumes aus N Daten ist für $N \to \infty$ asymptotisch äquivalent zu

$$l_N = (N+1)(\alpha + 1 - R^*(-1)) -$$

$$\frac{N}{Q_\infty} \cdot \delta_1(\log_2 N) - \frac{1}{Q_\infty} \delta_2(\log_2 N) + \mathcal{O}\left(\frac{1}{N}\right).$$

Hierbei sind $\delta_1(x)$ und $\delta_2(x)$ periodische Funktionen von Periode 1 mit kleiner Amplitude und Mittelwert 0; ihre Fourierentwicklungen sind

$$\delta_1(x) = \frac{1}{L} \sum_{k \neq 0} \chi_k \Gamma(-1-\chi_k) \cdot e^{2k\pi i x}$$

bzw.

$$\delta_2(x) = \frac{1}{L} \sum_{k \neq 0} \left(1 - \frac{\chi_k}{2}\right) \Gamma(1-\chi_k) \cdot e^{2k\pi i x}. \qquad \square$$

5. Varianz der Anzahl der internen Endknoten

In diesem Abschnitt soll nun die Varianz des gesuchten Parameters vermöge (3.4), d. h. durch die Formel

$$F''_N(1) + F'_N(1) - (F'_N(1))^2 \tag{5.1}$$

berechnet werden. Dazu muß der Ausdruck $F''_N(1)$ bestimmt werden, den wir abkürzend mit

$$w_N := F''_N(1) \tag{5.2}$$

bezeichnen. Aus der Rekursion (4.1) erhält man sofort durch zweimaliges Ableiten

$$w_{N+1} = 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} w_k + 2^{1-N} \sum_{k=0}^{N} \binom{N}{k} l_k\, l_{N-k},\ N \geqq 0,\ w_0 = 0. \tag{5.3}$$

Wie im 4. Abschnitt verwenden wir exponentielle erzeugende Funktionen:

$$W(z) := \sum_{N \geqq 0} w_N \frac{z^N}{N!} \quad \text{bzw.} \quad \hat{W}(z) = e^{-z} W(z) = \sum_{N \geqq 0} \hat{w}_N \frac{z^N}{N!}. \tag{5.4}$$

Damit läßt sich (5.3) schreiben als

$$W'(z) = 2\, e^{z/2}\, W\left(\frac{z}{2}\right) + 2\, L^2\left(\frac{z}{2}\right)$$

bzw. (5.5)

$$\hat{W}'(z) + \hat{W}(z) = 2\, \hat{W}\left(\frac{z}{2}\right) + 2\, \hat{L}^2\left(\frac{z}{2}\right).$$

Man kann nun die Koeffizienten ablesen:

$$\hat{w}_{N+1} = -(1 - 2^{1-N})\, \hat{w}_N + 2^{1-N} (-1)^N \sum_{k=2}^{N-2} \binom{N}{k} R_{k-2} R_{N-2-k} +$$

$$+ 2^{2-N} (-1)^N \cdot N \cdot R_{N-3},\ N \geqq 3,\ \hat{w}_0 = \hat{w}_1 = \hat{w}_2 = 0. \tag{5.6}$$

Wir kürzen ab:

$$\sigma_N := \sum_{k=2}^{N-2} \binom{N}{k} R_{k-2} R_{N-2-k}. \tag{5.7}$$

Die Rekursion (5.6) wird nun als

$$\frac{\hat{w}_{N+1}}{(-1)^N Q_{N-1}} = \frac{\hat{w}_N}{(-1)^{N-1} Q_{N-2}} + \frac{N \cdot 2^{2-N} R_{N-3}}{Q_{N-1}} + \frac{2^{1-N}}{Q_{N-1}} \sigma_N, \quad N \geqq 3,$$

$$\hat{w}_0 = \hat{w}_1 = \hat{w}_2 = 0, \ \hat{w}_3 = 1 \tag{5.8}$$

geschrieben und durch Summation gelöst:

$$\hat{w}_N = (-1)^{N-1} Q_{N-2} \cdot \left\{ 2 + \sum_{j=3}^{N-1} \frac{j \cdot 2^{2-j} R_{j-3}}{Q_{j-1}} + \sum_{j=4}^{N-1} \frac{2^{1-j}}{Q_{j-1}} \sigma_j \right\} \text{ für } N \geqq 4. \tag{5.9}$$

Man gewinnt dann w_N als

$$w_N = \binom{N}{3} + \sum_{k=4}^{N} \binom{N}{k} \hat{w}_k. \tag{5.10}$$

Zur asymptotischen Auswertung von w_N benötigen wir eine Fortsetzung von $\hat{w}_N$, welche auch für komplexe Werte sinnvoll ist. Dieses Problem wird nun in der Folge behandelt.

Die erste Schwierigkeit ist, daß die Summe

$$\sum_{j=4}^{N-1} 2^{-j} \sigma_j, \tag{5.11}$$

welche in (5.9) im wesentlichen auftritt, für $N \to \infty$ nicht konvergiert. (Man beachte $Q_N \sim Q_\infty$.) Man definiert deshalb den wesentlich kleineren Ausdruck

$$\sigma_N^* = \sum_{k=2}^{N-2} \binom{N}{k} R_{k-2}^* R_{N-2-k}^* \tag{5.12}$$

und rechnet vermöge (4.11) um:

$$\sigma_N = \sigma_N^* + Y_N + X_N \tag{5.13}$$

mit

$$Y_N = 2 \sum_{k=2}^{N-2} \binom{N}{k} R_{k-2}^* (N-k-1-\alpha)$$

und

$$X_N = \sum_{k=2}^{N-2} \binom{N}{k} (k-1-\alpha)(N-k-1-\alpha).$$

Nach mühsamer, aber elementarer Rechnung erhält man

$$\begin{aligned} X_N = {} & 2^{N-2}\{N^2 - (5+4\alpha)N + 4(\alpha+1)^2\} + \\ & + 2\alpha N^2 + 2N(1-\alpha-\alpha^2) - 2(\alpha+1)^2. \end{aligned} \tag{5.14}$$

Der Ausdruck für $\hat{w}_N$ in (5.9) zerfällt nun in fünf Bestandteile:

$$\hat{w}_N = (-1)^N \{f_1(N) + f_2(N) + f_3(N) + f_4(N) + f_5(N)\} \tag{5.15}$$

mit

$$\begin{aligned} f_1(N) &= -Q_{N-2}\cdot 2, \\ f_2(N) &= -Q_{N-2}\sum_{j=3}^{N-1}\frac{j\,2^{2-j}R_{j-3}}{Q_{j-1}}, \\ f_3(N) &= -Q_{N-2}\sum_{j=4}^{N-1}\frac{2^{1-j}}{Q_{j-1}}X_j, \\ f_4(N) &= -Q_{N-2}\sum_{j=4}^{N-1}\frac{2^{1-j}}{Q_{j-1}}Y_j, \\ f_5(N) &= -Q_{N-2}\sum_{j=4}^{N-1}\frac{2^{1-j}}{Q_{j-1}}\sigma_j^*. \end{aligned}$$

Als alternative Notation wollen wir auch

$$g_i(N) = f_i(N)/Q_{N-2} \quad \text{für } i = 1, \ldots, 5 \tag{5.16}$$

verwenden. Es gilt nun, die Größen $f_i(N)$ (bzw. $g_i(N)$) für komplexe Argumente fortzusetzen.

Die Fortsetzung von $g_1(N) = -2$ ist trivial, so daß wir uns

$$g_2(N) = -\sum_{j=3}^{N-1}\frac{j\,2^{2-j}R_{j-3}}{Q_{j-1}} \tag{5.17}$$

zuwenden können. Da $R_j = \mathcal{O}(j)$ ist, konvertiert die Summe in (5.17):

$$\rho_2 := \sum_{j\geqq 3}\frac{j\,2^{2-j}R_{j-3}}{Q_{j-1}}. \tag{5.18}$$

Damit folgt

$$g_2(N) = -\rho_2 + \sum_{j \geqq N} \frac{j\,2^{2-j} R_{j-3}}{Q_{j-1}} = \\ = -\rho_2 + \sum_{j \geqq 0} \frac{(j+N)\,2^{2-j-N} R_{j+N-3}}{Q_{j+N-1}}. \tag{5.19}$$

Somit ist die Fortsetzung erzielt:

$$g_2(z) = -\rho_2 + \sum_{j \geqq 0} \frac{(j+z)\,2^{2-j-z} R_{j+z-3}}{Q_{j+z-1}}. \tag{5.20}$$

Dabei ist R_{j+z-3} als

$$R_{j+z-3} = j + z - 2 - \alpha + R^*(j + z - 3)$$

bzw. Q_{j+z-1} als

$$Q_{j+z-1} = Q_\infty / Q\,(2^{1-j-z})$$

zu verstehen, wobei $Q(z)$ die in (3.5) definierte Funktion ist.

$g_3(N)$ ist unangenehmer, weil die Konvergenz erst durch Subtraktion geeigneter Terme erzwungen werden muß.

$$\begin{aligned} g_3(N) = &- \sum_{j=4}^{N-1} \frac{2^{1-j}}{Q_{j-1}} \,[2^{j-2}\{j^2 - (5+4\alpha)j + 4(\alpha+1)^2\} + \\ &+ 2\alpha j^2 + 2j(1-\alpha-\alpha^2) - 2(\alpha+1)^2] = \\ = &-\frac{1}{2} \sum_{j=4}^{N-1} \frac{1}{Q_{j-1}} \{j^2 - (5+4\alpha)j + 4(\alpha+1)^2\} - \\ &- \sum_{j=4}^{N-1} \frac{2^{2-j}}{Q_{j-1}} \{\alpha j^2 + j(1-\alpha-\alpha^2) - (\alpha+1)^2\} = \\ = &- g_{3,1}(N) - g_{3,2}(N). \end{aligned} \tag{5.21}$$

$g_{3,2}(N)$ ist konvergent; unsere Aufmerksamkeit müssen wir also $g_{3,1}(N)$ widmen und die nichtkonvergenten Bestandteile herausziehen. Dazu benötigen wir Informationen über

$$R_N = Q_N \sum_{j=0}^{N} \frac{1}{Q_j}, \quad S_N = Q_N \sum_{j=0}^{N} \frac{j}{Q_j} \quad \text{und} \quad T_N = Q_N \sum_{j=0}^{N} \frac{j^2}{Q_j}. \tag{5.22}$$

R_N ist schon bekannt: wir haben

$$R_N = N + 1 - \alpha + R_N^* \tag{5.23}$$

mit

$$R^*(z) = \frac{1}{Q_\infty} \sum_{i \geqq 1} \frac{1}{2^{z+i} - 1}\, a_i \{z + i\},$$

wobei wir in Hinkunft die Abkürzung

$$a_i = \frac{(-1)^{i-1}}{2^{\binom{i}{2}} Q_{i-1}} \tag{5.24}$$

verwenden wollen.

Als Hilfsresultat benötigen wir die Entwicklung der Funktion $1/Q(x)$ um $x = 1$, die durch eine elementare Rechnung zu erhalten ist:

$$\frac{Q_\infty}{Q(x)} = 1 - \alpha(1 - x) + \frac{\alpha^2 + \beta}{2} \cdot (1 - x)^2 + \dots \tag{5.25}$$

mit

$$\alpha = \sum_{k \geqq 1} \frac{1}{2^k - 1} \quad \text{und} \quad \beta = \sum_{k \geqq 1} \frac{1}{(2^k - 1)^2}.$$

Nun haben wir

$$\frac{S_N}{Q_N} = [x^N] \frac{1}{1 - x} \sum_{j \geqq 0} \frac{j}{Q_j} x^j =$$

$$= [x^N] \frac{x}{1 - x} \frac{d}{dx} \frac{1}{1 - x} \frac{1}{Q(x)} =$$

$$= [x^N] \frac{x}{1 - x} \left\{ \frac{1}{(1 - x)^2 Q(x)} + \frac{1}{1 - x} \cdot \frac{1}{Q(x)} \cdot \sum_{k \geqq 1} \frac{1}{2^k - x} \right\}.$$

Eine elementare Rechnung zeigt

$$\sum_{k \geqq 1} \frac{1}{2^k - x} = \alpha - \beta(1 - x) + \gamma_1 (1 - x)^2 + \dots \tag{5.26}$$

mit

$$\gamma_1 = \sum_{k \geqq 1} \frac{1}{(2^k - 1)^3}.$$

Durch Einsetzen von (5.25) und (5.26) sieht man nun

$$\frac{S_N}{Q_N} = [x^N] \frac{1}{Q_\infty} \left\{ \frac{1}{(1-x)^3} - \frac{1}{(1-x)^2} - \frac{\alpha^2 + \beta}{2} \frac{1}{1-x} + \dots \right\} =$$

$$= \frac{1}{Q_\infty} \left\{ \binom{N+2}{2} - (N+1) - \frac{\alpha^2 + \beta}{2} + \dots \right\}; \tag{5.27}$$

man kann daher ansetzen

$$S_N = \binom{N+1}{2} - \frac{\alpha^2 + \beta}{2} + S_N^*, \tag{5.28}$$

wo S_N^* klein genug ist, um die Konvergenz zu gewährleisten. Natürlich braucht man für S_N^* einen Ausdruck, der auch für komplexe z sinnvoll ist.

Aus der Definition von S_N ersieht man

$$S_{N+1} = (N+1) + \left(1 - \frac{1}{2^{N+1}}\right) S_N \tag{5.29}$$

und daher

$$S_N^* = \frac{1}{2^{N+1} - 1} \left[\binom{N+1}{2} - \frac{\alpha^2 + \beta}{2} \right] + \frac{1}{1 - \frac{1}{2^{N+1}}} S_{N+1}^* =$$

$$= \sum_{j \geqq 0} \left[\binom{j+N+1}{2} - \frac{\alpha^2 + \beta}{2} \right] \frac{2^{-N-j-1}}{\left(1 - \frac{1}{2^{N+1}}\right) \dots \left(1 - \frac{1}{2^{N+j+1}}\right)}. \tag{5.30}$$

Eine elementare Rechnung zeigt nun [vgl. (4.16) und (4.17)]

$$\sum_{j \geqq 0} \frac{1}{2^j Q_j} = \frac{1}{Q_\infty}, \quad \sum_{j \geqq 0} \frac{j}{2^j Q_j} = \frac{\alpha}{Q_\infty}, \quad \sum_{j \geqq 0} \frac{j(j-1)}{2^j Q_j} = \frac{\alpha^2 + \beta}{Q_\infty}. \tag{5.31}$$

Mit der Technik der Partialbruchzerlegung [vgl. (4.13)] und (5.31) ergibt sich $S_N^* = S^*(z)$ mit

$$S^*(z) = \frac{1}{Q_\infty} \sum_{i \geqq 1} \frac{1}{2^{z+i}-1} \cdot \\ \cdot a_i \left\{ \frac{z^2 + z(2i - 1 + 2\alpha) + i^2 + (2\alpha - 1)i}{2} \right\}. \tag{5.32}$$

Die gleiche Technik liefert auch die Fortsetzung von T_N; aus Gründen der Kürze wird die Rechnung hier übergangen.

$$T_N = 2\binom{N+1}{3} - (\alpha\beta + \gamma_1) + T_N^* \tag{5.33}$$

mit $T_N^* = T^*(N)$ und

$$T^*(z) = \frac{1}{Q_\infty} \sum_{i \geqq 1} \frac{1}{2^{i+1}-1} \cdot \\ \cdot a_i \left\{ 2\binom{z+i}{3} + 2\alpha\binom{z+i}{2} + (\alpha^2 + \beta)(z+i) \right\}.$$

Nun haben wir endlich

$$Q_{n-2} \cdot g_{3,1}(N) =$$

$$= Q_{n-2} \frac{1}{2} \sum_{j=4}^{N-1} \frac{1}{Q_{j-1}} \{j^2 - (5+4\alpha)j + 4(\alpha+1)^2\} =$$

$$= \frac{N-1}{2} \left(\frac{N^2}{3} - N\left(\frac{19}{6} + 2\alpha\right) + 4\alpha^2 + 8\alpha + 5 \right) -$$

$$- \alpha^3 - \frac{5}{4}\alpha^2 + \frac{\alpha\beta}{2} + \frac{3\beta}{4} - \frac{\gamma_1}{2} +$$

$$+ \frac{1}{2Q_\infty} \sum_{i \geqq 1} \frac{a_i}{2^{N+i-2}-1} \left\{ 2\binom{N+i-2}{3} - \right.$$

$$\left. - (3+2\alpha)\binom{N+i-2}{2} + (\alpha^2 + \alpha + \beta)(N+i-2) \right\} +$$

$$+ \frac{2}{3} Q_{N-2} \cdot (7 + 5\alpha - 17\alpha^2)$$

und damit

$$\begin{aligned} g_{3,1}(z) = \frac{1}{Q_{z-2}} \left\{ \frac{z-1}{2} \left(\frac{z^2}{3} - z\left(\frac{19}{6} + 2\alpha\right) + 4\alpha^2 + 8\alpha + 5 \right) - \right. \\ - \alpha^3 - \frac{5}{4}\alpha^2 + \frac{\alpha\beta}{2} + \frac{3\beta}{4} - \frac{\gamma_1}{2} + \qquad (5.34) \\ + \frac{1}{2Q_\infty} \sum_{i \geqq 1} \frac{a_i}{2^{z+i-2}-1} \cdot \left\{ 2\binom{z+i-2}{3} - (3+2\alpha)\binom{z+i-2}{2} + \right. \\ \left. + (\alpha^2+\alpha+\beta)(z+i-2) \right\} + \frac{2}{3}(7 + 5\alpha - 17\alpha^2). \end{aligned}$$

Die Betrachtung von $g_{3,2}(N)$ ist zum Glück einfacher:

$$\begin{aligned} \sum_{j=0}^{N} \frac{2^{-j}}{Q_j} &= \frac{1}{Q_\infty} - \sum_{i \geqq 1} \frac{2^{-N-i}}{Q_{N+i}} \\ \sum_{j=0}^{N} \frac{j2^{-j}}{Q_j} &= \frac{\alpha}{Q_\infty} - \sum_{i \geqq 1} \frac{(N+i)\,2^{-N-i}}{Q_{N+i}} \qquad (5.35) \\ \sum_{j=0}^{N} \frac{j^2 2^{-j}}{Q_j} &= \frac{\alpha^2+\alpha+\beta}{Q_\infty} - \sum_{i \geqq 1} \frac{(N+i)^2 2^{-N-i}}{Q_{N+i}}. \end{aligned}$$

Auf allen rechten Seiten kann man N durch $z \in \mathbb{C}$ ersetzen. Damit ergibt sich

$$\begin{aligned} g_{3,2}(z) = 2\,Q_{z-2} \left\{ \frac{\alpha(\beta-1)}{Q_\infty} + \frac{23}{3}\alpha^2 - \frac{2}{3}\alpha - \frac{7}{3} - \right. \\ \left. - \sum_{i \geqq 1} \frac{2^{-z-i+2}}{Q_{z+i-2}} \{-2 + (z+i)[(z+i-4)\alpha + 1 + \alpha - \alpha^2)]\} \right\}. \qquad (5.36) \end{aligned}$$

Bei $g_4(N)$ treten keine Konvergenzschwierigkeiten auf, so daß die Erweiterung durch eine Routinebetrachtung erfolgen kann:

$$\begin{aligned} g_4(N) &= -\sum_{j=4}^{N-1} \frac{2^{1-j}}{Q_{j-1}} Y_j = \\ &= -\sum_{j \geqq 4} \frac{2^{1-j}}{Q_{j-1}} Y_j + \sum_{j \geqq N} \frac{2^{1-j}}{Q_{j-1}} Y_j. \qquad (5.37) \end{aligned}$$

Bezeichnet man mit ρ_4 die auftretende Konstante:

$$\rho_4 := \sum_{j \geqq 4} \frac{2^{1-j}}{Q_{j-1}} Y_j, \tag{5.38}$$

dann kann man schreiben

$$g_4(z) = -\rho_4 + \sum_{j \geqq 0} \frac{2^{1-j-z}}{Q_{j+z-1}} Y(j+z); \tag{5.39}$$

man beachte, daß $Y(z)$ wie folgt eingeführt werden kann:

$$\begin{aligned} Y(z) = 2 \sum_{k \geqq 2} \binom{z}{k} R^*_{k-2}(z-k-1-\alpha) + 2\alpha z R^*(z-3) + \\ + 2(\alpha+1) R^*(z-2). \end{aligned} \tag{5.40}$$

Auch bei $g_5(N)$ treten keine Konvergenzschwierigkeiten auf, die Probleme sind vielmehr durch die Faltung in σ^*_N gegeben. Wieder wird die Technik der Partialbruchzerlegung zum Ziel führen. Zunächst kann man wie üblich schreiben

$$g_5(z) = -\rho_5 + \sum_{j \geqq 0} \frac{2^{1-j-z}}{Q_{j+z-1}} \sigma^*(j+z), \tag{5.41}$$

mit

$$\rho_5 := \sum_{j \geqq 4} \frac{2^{1-j}}{Q_{j-1}} \sigma^*_j,$$

nur muß man noch überlegen, was $\sigma^*(z)$ bedeuten soll. Dies geschieht in den nächsten Zeilen. Nach Definition und (4.13) ist

$$\sigma^*_N = \sum_{k=2}^{N-2} \binom{N}{k} \frac{1}{Q^2_\infty} \sum_{i,j \geqq 1} a_i a_j \frac{k-2+i}{2^{k-2+i}-1} \cdot \frac{N-k-2+j}{2^{N-k-2+j}-1}. \tag{5.42}$$

Wir schreiben

$$\begin{aligned} &\frac{1}{2^{k-2+i}-1} \cdot \frac{1}{2^{N-k-2+j}-1} = \\ &= \frac{1}{2^{N+i+j-4}-1} \left\{ 1 + \frac{1}{2^{k+i-2}-1} + \frac{1}{2^{N-k+j-2}-1} \right\} \end{aligned} \tag{5.43}$$

und setzen dies in (5.42) ein, was entsprechend den Ausdrücken in der geschweiften Klammer drei Reihen ergibt. In der dritten vertauschen wir i mit j und k mit $N-k$, was den Wert unverändert läßt, und fassen zusammen:

$$\begin{aligned}\sigma_N^* = \frac{1}{Q_\infty^2} \sum_{i,j \geqq 1} a_i a_j \frac{1}{2^{N+i+j-4}-1} \sum_{k=2}^{N-2} \binom{N}{k} \cdot \\ \cdot (k-2+i)(N-k-2+j) \cdot \left\{1 + \frac{2}{2^{k+i-2}-1}\right\}.\end{aligned} \tag{5.44}$$

Eine elementare Rechnung ergibt

$$\begin{aligned}\sum_{k=2}^{N-2} \binom{N}{k} (k-2+i)(N-k-2+j) = \\ = 2^{N-2}[N^2 + N(2i+2j-9) + 4(i-2)(j-2)] + \\ + [N^2(2-i-j) + N(-2+3i+3j-2ij) - 2(i-2)(j-2)].\end{aligned} \tag{5.45}$$

Die andere Summe (über k) kann durch Abzug zweier Terme von 2 bis unendlich erstreckt werden. Dies ergibt schließlich die gewünschte Fortsetzung:

$$\begin{aligned}\sigma^*(z) = \frac{1}{Q_\infty^2} \sum_{i,j \geqq 1} a_i a_j \frac{1}{2^{z+i+j-4}-1} \cdot \\ \cdot \Big\{2^{z-2}[z^2 + z(2i+2j-9) + 4(i-2)(j-2)] + \\ + [z^2(2-i-j) + z(-2+3i+3j-2ij) - \\ - 2(i-2)(j-2)] + 2 \sum_{k \geqq 2} (k-2+i)(z-k-2+j) \cdot \\ \cdot \binom{z}{k} \frac{1}{2^{k+i-2}-1} - 2(z-3+i)z(j-1)\frac{1}{2^{i+z-3}-1} - \\ - 2(z-2+i)(j-2)\frac{1}{2^{i+z-2}-1}\Big\}.\end{aligned} \tag{5.46}$$

Nachdem nun das Problem der Fortsetzungen abgeschlossen ist, können wir uns der asymptotischen Auswertung von w_N zuwenden. Wir wissen gemäß unserer einleitenden Ausführungen, daß

$$w_N \sim \binom{N}{3} + \sum \operatorname{res}\big([N;z]\{f_1(z) + \ldots + f_5(z)\}\big); \tag{5.47}$$

die Summe ist dabei über alle Pole mit $\operatorname{Re} z < 4$ zu erstrecken. Dies sind $3, 2+\chi_k, 1+\chi_k$ usw. ($k \in \mathbb{Z}$). Wir haben für $z \to 3$

$$[N; z] \sim \frac{1}{z-3}\binom{N}{3}$$

und $f_1(z) \sim -1$, während $f_i(z) \sim 0$ für $i = 2, 3, 4, 5$. Also hebt sich der Extraterm $\binom{N}{3}$ in (5.47) fort.

Wir schreiten nun zu $z \to 2$.

$$[N; z] \sim -\binom{N}{2}\frac{1}{z-2};$$

$$f_1(z) \sim -2$$

$$g_2(z) \sim \frac{2 \cdot 2^0 R_{-1}}{Q_1} = 4\,(-\alpha + R^*(-1))$$

$$f_2(z) \sim Q_0 \cdot g_2(z) \sim -4\alpha + 4R^*(-1)$$

$$\begin{aligned} g_{3,1}(z) \sim \frac{1}{Q_0}\Bigg\{&\frac{1}{2}\left(\frac{4}{3} - \frac{19}{3} - 4\alpha + 4\alpha^2 + 8\alpha + 5\right) - \\ &-\alpha^3 - \frac{5}{4}\alpha^2 + \frac{\alpha\beta}{2} + \frac{3\beta}{4} - \frac{\gamma_1}{2} + \\ &+ \frac{1}{2Q_\infty}\sum_{i \geqq 1}\frac{a_i}{2^i - 1}\left\{2\binom{i}{3} - (3 + 2\alpha)\binom{i}{2} + (\alpha^2 + \alpha + \beta)\,i\right\} + \\ &+ \frac{2}{3}(7 + 5\alpha - 17\alpha^2)\Bigg\} = -\frac{28}{3}\alpha^2 + \frac{16}{3}\alpha + \frac{14}{3}. \end{aligned}$$

Dabei sind folgende Formeln verwendet worden, welche man aus (5.23), (5.32) und (5.33) erhält:

$$\frac{1}{Q_\infty}\sum_{i \geqq 1}\frac{a_i}{2^i - 1}\binom{i}{m} = \begin{cases} \alpha & m = 1 \\ \dfrac{\beta - \alpha^2}{2} & m = 2 \\ \dfrac{\gamma_1 - 2\alpha\beta - \alpha^3}{2} & m = 3. \end{cases} \tag{5.48}$$

$$g_{3,2}(z) \sim \frac{34}{3}\alpha^2 - \frac{16}{3}\alpha - \frac{14}{3}. \tag{5.49}$$

$$g_4(z) \sim Y_2 = 4\alpha R^*(-1) \tag{5.50}$$

$$g_5(z) \sim \sigma^*(2).$$

Bei der Berechnung gemäß Formel (5.46) muß der Term für $i = j = 1$ gesondert betrachtet werden; er heiße vorübergehend $\sigma^*_{1,1}(2)$. Nach längerer elementarer Rechnung ergibt sich

$$\sigma^*_{1,1}(2) = \frac{1}{Q^2_\infty}\left[-6 + \frac{2}{L} - \frac{4}{L}\sum_{k \geqq 2}\frac{(-1)^k k}{(k+1)(k-1)(2^k-1)}\right]. \tag{5.51}$$

Der Beitrag $\sigma^*_{\geqq 2,1}(2)$ für $i \geqq 2$ und $j = 1$ ist

$$\sigma^*_{\geqq 2,1}(2) = 0. \tag{5.52}$$

Der Beitrag $\sigma^*_{1,\geqq 2}(2)$ für $i = 1$ und $j \geqq 2$ ist nach (4.13)

$$\begin{aligned}\sigma^*_{1,\geqq 2}(2) &= \frac{1}{Q^2_\infty}\cdot\frac{4}{L}\sum_{j \geqq 2}\frac{a_j}{2^{j-1}-1}(j-1) = \\ &= \frac{-4}{LQ_\infty}\left[R^*(-1) - \frac{1}{LQ_\infty}\right].\end{aligned} \tag{5.53}$$

Schließlich verbleibt noch für $i, j \geqq 2$:

$$\sigma^*_{\geqq 2,\geqq 2}(2) = -2\left[R^*(-1) - \frac{1}{LQ_\infty}\right]^2. \tag{5.54}$$

Wir fassen zusammen:

$$\begin{aligned}&g_1(2) + \ldots + g_5(2) \sim -2\cdot(\alpha + 1 - R^*(-1))^2 + \\ &+\frac{1}{Q^2_\infty}\left[-6 + \frac{2}{L} + \frac{2}{L^2} - \frac{4}{L}\sum_{k \geqq 2}\frac{(-1)^k k}{(k+1)(k-1)(2^k-1)}\right].\end{aligned} \tag{5.55}$$

Nun können wir den Koeffizienten C von N^2 in der gesuchten Varianz bestimmen, wobei wir

$$[N; z] \sim -\binom{N}{2}\frac{1}{z-2} \text{ für } z \to 2 \tag{5.56}$$

benützen. Er ist

$$C = [\alpha + 1 - R^*(-1)]^2 -$$

$$-\frac{1}{Q_\infty^2}\left[-3 + \frac{1}{L} + \frac{1}{L^2} - \frac{2}{L}\sum_{k\geqq 2}\frac{(-1)^k k}{(k+1)(k-1)(2^k-1)}\right] - \tag{5.57}$$

$$-[\alpha + 1 - R^*(-1)]^2 - \frac{1}{Q_\infty^2}[\delta_1^2]_0.$$

Wegen (2.18) haben wir

$$C = 0.$$

Wir können nun bereits unseren Hauptsatz formulieren:

Satz 2: Die Varianz der Anzahl der internen Endknoten eines digitalen Suchbaumes aus N Daten ist für $N \to \infty$ asymptotisch äquivalent zu

$$A \cdot N + N\,\delta_3(\log_2 N). \tag{5.58}$$

Dabei ist $\delta_3(x)$ eine periodische Funktion von Periode 1, kleiner Amplitude und Mittelwert 0; ihre Fourierkoeffizienten könnten im Prinzip angegeben werden.

$$\begin{aligned} A = &+\frac{1}{L}(g_2'(1) + \ldots + g_5'(1)) - 2(\alpha + 1 - R^*(-1))^2 + \\ &+ (\alpha + 1 - R^*(-1)) - \frac{1}{Q_\infty^2}[\delta_1^2]_0 - \frac{1}{Q_\infty^2}[\delta_1\delta_2]_0, \end{aligned} \tag{5.59}$$

wobei g_2, g_3, g_4 resp. g_5, die in (5.20), (5.34) und (5.36), (5.39) resp. (5.41) eingeführten Funktionen sind.

Beweis: Die Residuen bei $z = 2 + \chi_k$ geben Anlaß zu einer periodischen Funktion $\delta_4(x)$. Die Varianz besitzt die asymptotische Entwicklung

$$N^2\,\delta_4(\log_2 N) + \ldots$$

Wir behaupten, daß $\delta_4(x)$ identisch verschwindet. Im anderen Fall wäre $\delta_4(x) < -\varepsilon$ in einem Intervall $[a, b]$ für ein passendes $\varepsilon > 0$ (Mittelwert 0 und Stetigkeit).

Da $\log_2 N$ modulo 1 dicht ist, wäre die Varianz für unendlich viele N negativ — ein offensichtlicher Widerspruch.

Das Verschwinden von $\delta_4(x)$ führt auch zu unendlich vielen Identitäten, da die Fourierkoeffizienten identisch verschwinden müssen. In [5] sind solche explizit angegeben.

Es bleibt nur noch zu zeigen, daß bei $z = 1$ ein *einfacher* Pol vorliegt. Dazu ist zu zeigen, daß

$$g_1(1) + \ldots + g_5(1) = 0 \tag{5.60}$$

ist. Nun:

$$\begin{aligned}
g_1(1) &= -2; \\
g_2(1) &= 2 - 2\alpha + 2R^*(-1); \\
g_3(1) &= -2\alpha - 4\alpha^2; \\
g_4(1) &= 4\alpha(\alpha+1) + 2(1+2\alpha)R^*(-1); \\
g_5(1) &= \sigma^*(1) + \sigma^*(2),
\end{aligned}$$

und da

$$\sigma^*(1) = -\sigma^*(2) - 4(\alpha+1)R^*(-1)$$

gilt, folgt

$$g_5(1) = -4(\alpha+1)R^*(-1).$$

Insgesamt ergibt sich also (5.60).

Das Residuum von

$$[N; z](f_1(z) + \ldots + f_5(z))$$

an der Stelle $z = 1$ ist

$$\frac{N}{L}(g_1'(1) + \ldots + g_5'(1)). \tag{5.61}$$

Das ist aber nicht der einzige Beitrag zum linearen Term. Vom Residuum bei $z = 2$ kommt dazu:

$$\frac{N}{2}(g_1(2) + \ldots + g_5(2)). \tag{5.62}$$

Von l_N kommt

$$N(\alpha + 1 - R^*(-1)); \tag{5.63}$$

von $-l_N^2$ kommt

$$-N\left[(\alpha+1-R^*(-1))^2+\frac{1}{Q_\infty^2}[\delta_1\delta_2]_0\right], \tag{5.64}$$

was zur Konstanten A führt. Eine genauere Berechnung von A dürfte hoffnungslos sein.

Literatur

[1] Apostol, T. M.: Modular Functions and Dirichlet Series in Number Theory. Springer, New York 1976.

[2] Berndt, B. C.: Modular transformations and generalizations of several formulae of Ramanujan. Rocky Mountain J. Math. **7** (1977), 147—189.

[3] Flajolet, P., and R. Sedgewick: Digital search trees revisited. SIAM J. Comput. **15** (1986), 748—767.

[4] Jacquet, P., and M. Régnier: Normal limiting distributions of the size of tries, in: Performance '87. Proceedings of the 13th International Symposium on Computer Performance, P.-J. Courtois and G. Latouche, editors, Elsevier North Holland 1987, 209—223.

[5] Kirschenhofer, P., and H. Prodinger: On some Applications of Formulae of Ramanujan in the Analysis of Algorithms. Preprint 1987.

[6] Kirschenhofer, P., und H. Prodinger: Asymptotische Untersuchungen über charakteristische Parameter von Suchbäumen, in: Zahlentheoretische Analysis 2, E. Hlawka, ed., Lecture Notes in Mathematics **1262**, 93—107, Springer, Berlin 1987.

[7] Kirschenhofer, P., and H. Prodinger: Further results on digital search trees. Theor. Comput. Sci. **58** (1988), 143—154.

[8] Kirschenhofer, P., H. Prodinger und J. Schoissengeier: Zur Auswertung gewisser Reihen mit Hilfe modularer Funktionen, in: Zahlentheoretische Analysis 2, E. Hlawka, ed., Lecture Notes in Mathematics **1262**, 108—110, Springer, Berlin 1987.

[9] Kirschenhofer, P., H. Prodinger, and W. Szpankowski: On the balance property of Patricia tries: External path length viewpoint. Theoret. Comput. Sci., to appear.

[10] Kirschenhofer, P., H. Prodinger, and W. Szpankowski: On the variance of the external path length in a binary digital trie. Discrete Appl. Math., to appear.

[11] Knuth, D. E.: The art of computer programming, Vol. 3: "Sorting and searching". Addison Wesley, Reading Mass. 1973.

[12] Louchard, G.: Exact and asymptotic distributions in digital and binary search trees. R.A.I.R.O. Theoretical Informatics and Applications **21** (1987), 479—496.

[13] Nörlund, N. E.: Vorlesungen über Differenzenrechnung. Chelsea, New York 1954.

[14] Ramanujan, S.: Notebooks of Srinivasa Ramanujan (2 volumes). Tata Institute of Fundamental Research, Bombay 1957.

[15] Schmid, U.: Analyse von Collision-Resolution Algorithmen in Random-Access-Systemen mit dominanten Übertragungskanälen. Dissertation TU Wien, 1986.

[16] Siegel, C. L.: A simple proof of $\eta(-1/\tau)=\eta(\tau)\sqrt{\tau/i}$. Mathematika **1** (1954), 4.