ASYMPTOTISCHE UNTERSUCHUNGEN OBER CHARAKTERISTISCHE PARAMETER VON SUCHBÄUMEN

P.Kirschenhofer und H.Prodinger

Abstract. In Computer Science there are several algorithms which build search trees using digital properties of the search keys. An important parameter in the analysis of these data structures is the so-called pathlength. The average case analysis involves interesting mathematical methods from the calculus of finite differences and from asymptotic analysis. Some of the appearing constants may be evaluated using properties of modular functions.

1. EINLEITUNG

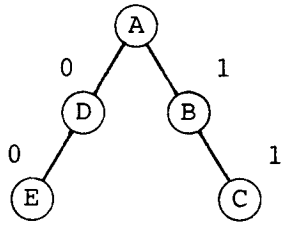
Eine wichtige Klasse von Algorithmen in der Informatik befaßt sich mit dem Abspeichern und Aufsuchen von Daten in geeigneten Datenstrukturen. Verwendung finden hiebei u.a. sogenannte Digitale Suchbäume, "Tries"(von "information retrieval"), "Patricia Tries"(von "Practical Algorithm To Retrieve Information Coded In Alphanumeric"). Wir werden im weiteren eine kurze Beschreibung dieser Datenstrukturen vornehmen, verweisen jedoch für eine ausführliche Darstellung, insbesondere der informatischen Aspekte, auf [4] und [5]. Der Schwerpunkt dieser Arbeit liegt auf der Analyse charakteristischer Parameter dieser Datenstrukturen mit Hilfe von Methoden der kombinatorischen und asymptotischen Analysis. (Ein Spezialfall der hier erzielten Resultate wurde in der Arbeit [3] untersucht.)

Wir betrachten im folgenden *Schlüssel* bzw. deren interne M-äre Darstellung, d.h. eine (potentiell unendliche) Folge von "Ziffern" $0,1,\ldots,M-1$. Als zu Grunde liegendes wahrscheinlichkeitstheoretisches Modell nehmen wir das der Gleichverteilung.

Unter einem *Digitalen Suchbaum* verstehen wir einen M-ären Baum, dessen Knoten die gegebenen Schlüssel $A_1,\ldots,A_N$ beinhalten, und der folgendermaßen aufgebaut wird: A_1 wird in der Wurzel abgespeichert; seien $A_1,\ldots,A_{i-1}$ schon abgespeichert, dann wird der Platz für $A_i = b_1b_2b_3\ldots$ so gefunden: Man betrachtet den b_1-ten Teilbaum der Wurzel, dann den b_2-ten Teilbaum dessen Wurzel, usw., bis zum erstenmal ein noch nicht durch einen Schlüssel besetzter Knoten auftritt.

Beispiel für M=2:

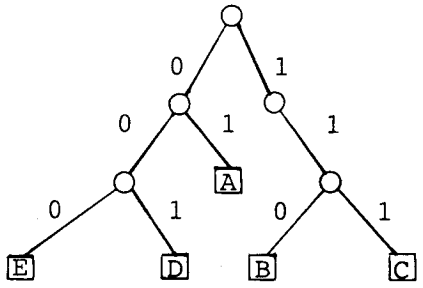
A : 010...
B : 110...
C : 111...
D : 001...
E : 000...



Man beachte, daß die Reihenfolge, in der die Schlüssel eingetragen werden, relevant ist!

Für die Konstruktion der M-ären *Tries* verwendet man im wesentlichen dieselbe Idee, nur werden die Daten nur in den Blättern (Endknoten) abgespeichert:

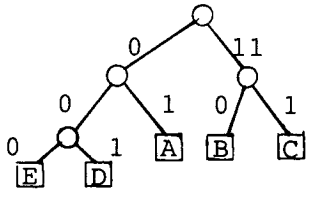
Beispiel für M=2:



Hier ist die relative Ordnung der Schlüssel irrelevant.

Patricia Tries entstehen aus Tries durch Komprimierung: Falls ein interner Knoten nur einen einzigen Nachfolger hat, wird er und die nachfolgende Kante weggelassen; statt dessen wird die Information der wegfallenden Kante (bzw.die zu überspringenden Ziffern) an passender Stelle gemerkt. Es sei angemerkt, daß Patricia Tries in der Literatur nur für den Fall M=2 vorkommen; die gegenständliche Verallgemeinerung erscheint die plausibelste zu sein.

Beispiel für M=2 (Fortsetzung von oben):



BEMERKUNG. Im weiteren bezeichnen wir mit $[z^k]f(z)$ den Koeffizienten von z^k in der Potenzreihe $f(z)$.

2. ERZEUGENDE FUNKTIONEN UND ERWARTUNGSWERTE

Ein wichtiger Parameter in der Anwendung der betrachteten Datenstrukturen in der Informatik ist die *Anzahl der Suchschritte* beim erfolgreichen Aufsuchen eines Schlüssels. Diese gewinnt man auf dem Umweg über die Betrachtung der *internen* bzw. *externen Pfadlänge* der zu Grunde liegenden Bäume:

Für Tries bzw. Patricia Tries studiert man die externe Pfadlänge; das ist die Summe der Abstände aller Endknoten von der Wurzel (gemessen in Kanten). Für Digitale Suchbäume betrachtet man statt dessen die interne Pfadlänge: das ist die Summe der Abstände aller Knoten von der Wurzel.

Die (kumulierte) Anzahl der Suchschritte beim erfolgreichen Aufsuchen aller Schlüssel eines festen Baumes ist dann die jeweils relevante Pfadlänge. Ist der betrachtete Baum aus N Daten entstanden, so kann die durch N dividierte obige Anzahl als mittlere Anzahl der Suchschritte bei erfolgreicher Suche gedeutet werden.

Demnach gilt: Der Erwartungswert für die Anzahl der Suchschritte bei erfolgreicher Suche in irgendeinem aus N Daten entstandenen Baum ist gleich dem 1/N-fachen des Erwartungswerts der Pfadlänge aller derartigen Bäume. Wir betrachten daher im weiteren nur den letzteren Begriff.

Im folgenden bezeichne $h_N^{[T]}(z)$ die erzeugende Funktion, deren Koeffizienten $[z^k]h_N^{[T]}(z)$ den Erwartungswert der Anzahl der externen Knoten auf dem Niveau k angibt für die Familie aller Tries, die aus N zufälligen Schlüsseln gebildet sind. (Die Wurzel steht auf Niveau 0.)

$h_N^{[P]}(z)$ bezeichne die analoge erzeugende Funktion für Patricia Tries.

$h_N^{[D]}(z)$ steht für Digitale Suchbäume, wobei allerdings an Stelle der externen Knoten die internen gezählt werden.

(Wenn keine Verwechslungsgefahr besteht, werden wir den oberen Index fortlassen.)

$h_N'(1)$ gibt dann den Erwartungswert der relevanten Pfadlänge an.

Weiters gibt der Ausdruck

$$h_N''(1) + h_N'(1) - \frac{1}{N}(h_N'(1))^2$$

die *Varianz* der relevanten Pfadlänge in folgendem Modell an: Anstelle aller Bäume aus N Daten betrachten wir jetzt nur *abstrakte Durch-*

schnittsbäume in deren k-tem Niveau $[z^k]h_N(z)$ Knoten (intern bzw. extern) vorhanden sind. Das Studium dieser Größe wird dem Kapitel 3 vorbehalten sein.

Aus der Definition der Tries ergibt sich folgende Rekursion:

$$h_N^{[T]}(z) = M^{-N} \sum_{\Sigma k_j = N} \binom{N}{k_1,\ldots,k_M} z \left(h_{k_1}^{[T]}(z)+\ldots+h_{k_M}^{[T]}(z)\right)$$

$$= M^{1-N} \sum_{k=0}^{N} z\binom{N}{k} h_k^{[T]}(z)(M-1)^{N-k}, \quad N\geq 2, \quad h_0^{[T]}(z)=0, h_1^{[T]}(z)=1. \tag{2.1}$$

Die Rekursion für Patricia Tries ist ähnlich:

$$h_N^{[P]}(z) = M^{1-N} \sum_{k=0}^{N} z\binom{N}{k} h_k^{[P]}(M-1)^{N-k} - M^{1-N}(z-1)h_N^{[P]}(z), \quad N\geq 2,$$

$$h_0^{[P]}(z) = h_1^{[P]}(z) = 1. \tag{2.2}$$

Der Unterschied ergibt sich aus dem Falle, in welchem die Wurzel im zugrunde liegenden Trie genau einen Nachfolger hat.

Für M-äre Digitale Suchbäume erhalten wir die (für M>2 neue) Formel

$$h_N^{[D]}(z) = \sum_{k\geq 0} \binom{N}{k+1} (-1)^k \prod_{0\leq j<k} \left(1-\frac{z}{M^j}\right). \tag{2.3}$$

Beweis. Seien a(z) und b(z) die erzeugenden Funktionen für die Anzahlen der internen bzw. der externen Knoten auf Niveau k≥0. Dann gilt für jeden festen Baum (und damit auch für jede Familie von Bäumen) offensichtlich der folgende Zusammenhang:

$$b(z) = 1+(Mz-1)a(z) \tag{2.4}$$

und daher

$$a(z)+b\left(\frac{z}{M}\right) = a(z)+1+(z-1)a\left(\frac{z}{M}\right). \tag{2.5}$$

Sei nun $\tilde{h}_N(z)$ das Analogon zu $h_N(z)$ welches die externen Knoten zählt; dann gilt wegen (2.4)

$$\tilde{h}_N(z) = 1 + (Mz-1)h_N(z). \tag{2.6}$$

Da im k-ten Niveau ein externer Knoten mit Wahrscheinlichkeit M^{-k} auftritt, erhalten wir die folgende Rekursion:

$$h_{N+1}(z) = h_N(z) + \tilde{h}_N\left(\frac{z}{M}\right).$$

Aus (2.6) ergibt sich damit

$$h_{N+1}(z) = h_N(z)+1+(z-1)h_N\left(\frac{z}{M}\right), N\geq 0, h_0(z)=0$$

bzw.

(2.7) $\tilde{h}_{N+1}(z) = \tilde{h}_N(z)+(Mz-1)\tilde{h}_N\left(\frac{z}{M}\right)$, $N\geq 0$, $\tilde{h}_0(z)=1$.

Durch Iteration von (2.7) erhält man

$$\tilde{h}_N(z) = \sum_{k\geq 0}\binom{N}{k}\prod_{0\leq j<k}(M^{1-j}z-1),$$

woraus sich wegen (2.6) die angekündigte Formel (2.3) ergibt. □

Gemäß den einleitenden Bemerkungen wenden wir uns nun dem Studium der Größe $h'_N(1)$, welche den Erwartungswert der relevanten Pfadlänge angibt, zu.

SATZ 1. [4]
Der Erwartungswert der externen Pfadlänge eines M-ären Tries aus N Daten mit zufälligen Schlüsseln ist für $N\to\infty$.

$$N\left(\log_M N + \frac{\gamma}{\log M} + \frac{1}{2} + \delta^{[T]}(\log_M N)\right) + O(1),$$

wobei γ die Eulersche Konstante ist und $\delta^{[T]}(x)$ eine periodische Funktion mit Periode 1 und sehr kleiner Amplitude für kleine Werte von M:

$$\delta^{[T]}(x) = \frac{1}{\log M}\sum_{k\in\mathbb{Z},k\neq 0}\omega_k\cdot\Gamma(-\omega_k)e^{2k\pi ix} \text{ mit } \omega_k = 1+\frac{2k\pi i}{\log M}.$$ □

Der Beweis von Satz 1 wird von Knuth mit Hilfe der *Mellin-Transformation* geführt; wir wollen in dieser Arbeit jedoch eine andere, auf Rice zurückgehende, Beweistechnik verwenden. Diese hat sich bereits in [2] als vorteilhaft erwiesen.

SATZ 2. Der Erwartungswert der externen Pfadlänge eines M-ären Patricia Tries aus N Daten mit zufälligen Schlüsseln ist für $N\to\infty$

$$N\left(\log_M N + \Delta_M + \frac{\gamma}{\log M} - \frac{1}{2} + \delta^{[T]}(\log_M N+\Delta_M)\right) + O(1)$$

mit $\delta^{[T]}(x)$ aus Satz 1 und $\Delta_M = \log_M(M-1)$.

Beweis. Aus (2.2) ergibt sich wegen $h_N(1)=N$

$$h'_N(1) = N+M^{1-N}\sum_{k\geq 0}\binom{N}{k}h'_k(1)(M-1)^{N-k} - M^{1-N}\cdot N.$$

Wir setzen

$$R(z) = \sum_{N\geq 0} h'_N(1)\frac{z^N}{N!} \text{ sowie } V(z) = e^{-z}R(z) = \sum_{N\geq 0} v_N\frac{z^N}{N!}.$$

Dann erhalten wir

$$R(z) = z(e^z-1)+Me^{z(1-\frac{1}{M})}R\left(\frac{z}{M}\right) - M\frac{z}{M}(e^{z/M}-1)$$

bzw.

$$V(z) = z(1-e^{-z}) + M\cdot V\left(\frac{z}{M}\right) - z\left(e^{-z(1-\frac{1}{M})} - e^{-z}\right)$$
$$= z\left(1-e^{-z(1-\frac{1}{M})}\right) + M\cdot V\left(\frac{z}{M}\right)$$

und daher (N≥2)

$$v_N = -N(1-\tfrac{1}{M})^{N-1}\cdot(-1)^{N-1} + M^{1-N}v_N$$
$$= \frac{N(-1)^N(M-1)^{N-1}}{M^{N-1}-1} \; ; \; v_0 = v_1 = 0.$$

Damit ist (N≥2)

$$h_N'(1) = \sum_{k\geq 2}\binom{N}{k} v_k = \sum_{k\geq 2}\binom{N}{k}\frac{k\cdot(-1)^k(M-1)^{k-1}}{M^{k-1}-1} \; .$$

Zur asymptotischen Auswertung dieses Ausdrucks bedienen wir uns des folgenden Lemmas:

LEMMA 3.[6] Sei $\mathcal{C}$ eine Kurve, die die Punkte 2,...,N in $\mathbb{C}$ umschließt, und sei f(z) analytisch im Inneren und stetig am Rand von $\mathcal{C}$. Dann gilt:

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

mit der Abkürzung

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

Sei nun

$$f(z) = \frac{z(M-1)^{z-1}}{M^{z-1}-1} \; .$$

Durch Anwendung des Lemmas und Ausdehnung des Integrationswegs auf die linke Seite der Geraden Re z = 1 (vgl.[2] für technische Einzelheiten) erhalten wir:

$$h_N'(1) \sim \sum_{k\in\mathbb{Z}} \mathrm{Res}([N;z]f(z);z=\omega_k).$$

Zur Berechnung der Residuen müssen wir die lokalen Entwicklungen von [N;z]f(z) um die Pole $\omega_0 = 1$(zweifach) und ω_k, $k\neq 0$(einfach) bestimmen. Dazu verwenden wir:

Mit u = z-1, L = log M, L' = log(M-1) gilt:

$$[N;z]\sim \frac{N}{u}\left(1+u(H_{N-1}-1)+u^2(1-H_{N-1}+\tfrac{1}{2}H_{N-1}^2+\tfrac{1}{2}H_{N-1}^{(2)})\right)$$

(2.8) $H_N = \sum_{1\leq k\leq N}\frac{1}{k}$ und $H_N^{(2)} = \sum_{1\leq k\leq N}\frac{1}{k^2}$ bezeichnen *Harmonische Zahlen* 1. bzw. 2.Ordnung.

$$(2.8)\quad \begin{aligned} z &\sim 1+u \\ (M-1)^{z-1} &\sim 1+uL'+\frac{u^2}{2}L'^2 \\ \frac{1}{M^{z-1}-1} &\sim \frac{1}{Lu}\left(1-\frac{Lu}{2}+\frac{L^2}{12}u^2\right) \end{aligned}$$

Daher ist

$$\mathrm{Res}([N;z]f(z);z=1) = \frac{N}{L}\,(H_{N-1}+L'-\tfrac{L}{2}).$$

(Beim Studium der Varianz werden wir die Terme höherer Ordnung der obigen Entwicklungen verwenden.)

Mit $u = z-\omega_k$, $k\neq 0$ gilt

$$[N;z]\sim N^{\omega_k}(\Gamma(-\omega_k)+u(-\Gamma'(-\omega_k)+\Gamma(-\omega_k)\cdot\log N))$$

$$(2.9)\quad \begin{aligned} z &\sim \omega_k+u \\ (M-1)^{z-1} &\sim e^{\frac{L'}{L}\cdot 2k\pi i}(1+L'u) \\ \frac{1}{M^{z-1}-1} &\sim \frac{1}{Lu}-\frac{1}{2}. \end{aligned}$$

Daher ist für $k\neq 0$

$$\begin{aligned} \mathrm{Res}([N;z]f(z);z=\omega_k) &= N^{\omega_k}\Gamma(-\omega_k)\omega_k\, e^{\frac{L'}{L}2k\pi i}\cdot\frac{1}{L} \\ &= \frac{N}{L}\,\omega_k\Gamma(-\omega_k)\, e^{2k\pi i(\log_M N+\frac{L'}{L})}. \end{aligned}$$

□

BEMERKUNG. Im Falle $M=2$ gilt für die Erwartungswerte (exakt!)

$$E_N^{[P]} = E_N^{[T]}-N;$$

für allgemeines $M\geq 2$ gilt nach den Sätzen 1 und 2

$$E_N^{[P]} = \frac{1}{M-1}\,E_{N(M-1)}^{[T]} - N + O(1)$$

sowie

$$E_N^{[P]} = E_N^{[T]}-N(1-\Delta_M)-N(\delta^{[T]}(\log_M N)-\delta^{[T]}(\log_M N+\Delta_M)) + O(1).$$

SATZ 4. Der Erwartungswert der internen Pfadlänge eines M-ären Digitalen Suchbaums aus N Daten mit zufälligen Schlüsseln ist

$$N\left(\log_M N+\frac{\gamma-1}{\log M}+\frac{1}{2}-\alpha+\delta^{[D]}(\log_M N)\right)+O(N^{1/2})$$

mit

$$\alpha = \sum_{k\geq 1}\frac{1}{M^k-1} \quad\text{und}\quad \delta^{[D]}(x) = \frac{1}{\log M}\sum_{k\in\mathbb{Z},k\neq 0}\Gamma(-\omega_k)e^{2k\pi ix}.$$

Beweis. Aus (2.3) ergibt sich durch Differentiation

$$h'_N(1) = \sum_{k\geq 2} \binom{N}{k} (-1)^k \prod_{1\leq j\leq k-2} \left(1-\frac{1}{M^j}\right) ;$$

die weitere Beweisführung kann daher in völliger Analogie zum Fall M=2 in [2] erfolgen. □

3. VARIANZEN

Gemäß unserer einführenden Bemerkungen werden wir nun den Ausdruck

$$h''_N(1) + h'_N(1) - \frac{1}{N}\left(h'_N(1)\right)^2$$

für die 3 Datenstrukturen untersuchen.

SATZ 5. Die Varianz der externen Pfadlänge eines M-ären Tries aus N Daten mit zufälligen Schlüsseln ist für $N\to\infty$ asymptotisch gleich

$$N\left(\frac{1}{12} + \frac{\pi^2}{6\log^2 M} + \sigma^{[T]}(\log_M N)\right)$$

mit der periodischen Funktion

$$\sigma^{[T]}(x) = \frac{2}{\log^2 M} \sum_{k\neq 0} \left(\Gamma(-\omega_k) - \omega_k\Gamma'(-\omega_k) - \gamma\omega_k\Gamma(-\omega_k)\right)\cdot e^{2k\pi i x} - \left(\delta^{[T]}(x)\right)^2 .$$

Beweis. Wir verwenden die folgenden erzeugenden Funktionen:

$$R(z) = \sum_{N\geq 0} h'_N(1)\frac{z^N}{N!} ,$$

$$S(z) = \sum_{N\geq 0} h''_N(1)\frac{z^N}{N!} ,$$

$$V(z) = e^{-z}R(z) = \sum_{N\geq 0} v_N\frac{z^N}{N!} ,$$

$$W(z) = e^{-z}S(z) = \sum_{N\geq 0} w_N\frac{z^N}{N!} .$$

Aus der Rekursion (2.1) erhalten wir

$$h''_N(1) = M^{1-N}\sum_{k\geq 0}\binom{N}{k}(M-1)^{N-k}\left(h''_k(1)+2h'_k(1)\right),\quad N\geq 0$$

und daher

$$S(z) = M\cdot S\left(\frac{z}{M}\right)e^{z(1-\frac{1}{M})} + 2M\cdot R\left(\frac{z}{M}\right)e^{z(1-\frac{1}{M})}$$

bzw.

$$W(z) = M\cdot W\left(\frac{z}{M}\right) + 2M\cdot V\left(\frac{z}{M}\right) .$$

Da andererseits

$$V(z) = M\cdot V\left(\frac{z}{M}\right) + z(1-e^{-z})$$

ist, folgt

$$v_N = \frac{N\cdot(-1)^N}{1-M^{1-N}}\ ,\ N\geq 2,\ v_0 = v_1 = 0.$$

Also ergibt sich

$$w_N = \frac{2\cdot M^{1-N}\cdot N\cdot(-1)^N}{(1-M^{1-N})^2}\ ,\ N\geq 2,\ w_0 = w_1 = 0$$

und daher (N≥2)

$$h_N''(1) = \sum_{k\geq 2}\binom{N}{k}(-1)^k\ \frac{2\cdot M^{1-k}\cdot k}{(1-M^{1-k})^2}\ .$$

Nach Anwendung von Lemma 3 erhalten wir

$$h_N''(1) \sim \sum_{k\in\mathbb{Z}} \mathrm{Res}([N;z]f(z);z=\omega_k)$$

mit

$$f(z) = \frac{2\cdot M^{1-z}\cdot z}{(1-M^{1-z})^2}\ .$$

Wir haben die Pole $\omega_0=1$(dreifach) und $\omega_k = 1+\frac{2k\pi i}{\log M}$, $k\neq 0$(doppelt). Zur Berechnung der Residuen verwenden wir die lokalen Entwicklungen (2.8) und (2.9) und erhalten (mit L = log M)

$$h_N''(1) \sim \frac{N}{L^2}\,(H_{N-1}^2+H_{N-1}^{(2)}) - \frac{N}{6}$$

$$+\frac{2N}{L^2}\sum_{k\neq 0}\left(\Gamma(-\omega_k)+\omega_k(-\Gamma'(-\omega_k)+\Gamma(-\omega_k)\log N)\right)e^{2k\pi i\,\log_M N},$$

woraus sich das Ergebnis unter Anwendung des asymptotischen Äquivalents der Harmonischen Zahlen, sowie des asymptotischen Äquivalents von $h_N'(1)$ (vergleiche Satz 1), ergibt. □

SATZ 6. Die Varianz der externen Pfadlänge eines M-ären Patricia Tries aus N Daten mit zufälligen Schlüsseln ist für N→∞ asymptotisch gleich

$$N\left(\frac{1}{12}-\frac{2}{\log M}\cdot\mu+\frac{\pi^2}{6\log^2 M}+\sigma^{[P]}(\log_M N+\Delta_M)\right)$$

mit der periodischen Funktion

$$\sigma^{[P]}(x) = \sigma^{[T]}(x) - \frac{2}{\log M}\sum_{k\neq 0}\omega_k\Gamma(-\omega_k)\cdot(\Delta_M+\xi_k)e^{2k\pi ix},$$

wobei $\mu = \sum_{n\geq 1}\frac{(-1)^{n-1}}{n(M^n-1)}$, $\Delta_M = \log_M(M-1)$ und $\xi_k = \sum_{n\geq 1}\binom{\frac{2k\pi i}{\log M}}{n}\frac{1}{M^n-1}$.

Beweis. Wir verwenden wieder die (analogen) erzeugenden Funktionen und erhalten aus Rekursion (2.2)

$$W(z) = M\cdot W\left(\frac{z}{M}\right) + 2M\left(1-e^{-z(1-\frac{1}{M})}\right)V\left(\frac{z}{M}\right).$$

Aus dem Beweis von Satz 2 wissen wir

$$v_N = \frac{N(-1)^N(M-1)^{N-1}}{M^{N-1}-1}\ ,\ N\geq 2,\ v_0 = v_1 = 0,$$

sodaß ($N\geq 2$)

$$w_N = \frac{2N(-1)^N(M-1)^{N-1}}{(M^{N-1}-1)^2} - \frac{2N(-1)^N(M-1)^{N-1}}{M^{N-1}-1} \sum_{n\geq 1} \binom{N-1}{n} \frac{1}{M^n-1}\ .$$

Damit ist

$$h_N''(1) = \sum_{k\geq 2} \binom{N}{k} w_k,$$

und Lemma 3 kann angewendet werden mit

$$f(z) = \frac{2z(M-1)^{z-1}}{(M^{z-1}-1)^2} - \frac{2z(M-1)^{z-1}}{M^{z-1}-1} \sum_{n\geq 1} \binom{z-1}{n} \frac{1}{M^n-1}\ .$$

Wir verwenden wieder die lokalen Entwicklungen (2.8) und

$$\sum_{n\geq 1} \binom{z-1}{n} \frac{1}{M^n-1} \sim u\cdot \sum_{n\geq 1} \frac{(-1)^{n-1}}{n(M^n-1)} \text{ für } u = z-1 \to 0.$$

Damit ergibt sich

$\mathrm{Res}([N;z]f(z);z=1) =$

$$N\Big((\log_M N)^2 + 2\,\frac{\gamma-L+L'}{L}\log_M N + \frac{2\gamma L'+L'^2+\gamma^2+\frac{\pi^2}{6}}{L^2} - 2\,\frac{\gamma+L'}{L} + \frac{5}{6}$$

$$- \frac{2}{L}\cdot \sum_{n\geq 1} \frac{(-1)^{n-1}}{n(M^n-1)}\Big).$$

Nun verwenden wir die lokalen Entwicklungen (2.9) und ($z \to \omega_k, k\neq 0$)

$$\sum_{n\geq 1} \binom{z-1}{n} \frac{1}{M^n-1} \sim \sum_{n\geq 1} \binom{\frac{2k\pi i}{\log M}}{n} \frac{1}{M^n-1} =: \xi_k.$$

Damit erhalten wir

$\mathrm{Res}([N;z]f(z);z=\omega_k) =$

$$2N\left\{ \frac{1}{L^2}\Big(\Gamma(-\omega_k)+\omega_k(-\Gamma'(-\omega_k)+\Gamma(-\omega_k)\log N)\Big) \right.$$

$$\left. - \frac{1}{L}\,\omega_k\Gamma(-\omega_k)\cdot(1+\xi_k) \right\} e^{2k\pi i(\log_M N+\Delta_M)}$$

mit $\Delta_M = L'/L = \log_M(M-1)$.

Die Summe der Residuen ergibt nun das asymptotische Äquivalent für $h_N''(1)$; die Varianz $h_N''(1)+h_N'(1)-\frac{1}{N}(h_N'(1))^2$ ergibt sich nun durch Kombination mit Satz 2. □

SATZ 7. Die Varianz der internen Pfadlänge eines M-ären Digitalen Suchbaumes aus N Daten mit zufälligen Schlüsseln ist für $N \to \infty$ asymptotisch gleich

$$N\left(\frac{1}{12} + \frac{\pi^2}{6\log^2 M} + \frac{1}{\log^2 M} - \alpha - \beta + \sigma^{[D]}(\log_M N)\right)$$

mit

$$\alpha = \sum_{k\geq 1} \frac{1}{M^k-1}, \quad \beta = \sum_{k\geq 1} \frac{1}{(M^k-1)^2}$$

und

$$\sigma^{[D]}(x) = \frac{2}{\log^2 M} \sum_{k\neq 0} \left(-\Gamma'(-\omega_k)+(1-\gamma)\Gamma(-\omega_k)\right) e^{2k\pi i x} - \left(\delta^{[D]}(x)\right)^2.$$

Beweis. In (2.3) haben wir bewiesen:

$$h_N(z) = \sum_{k\geq 0} \binom{N}{k+1} (-1)^k \prod_{0\leq j<k} \left(1-\frac{z}{M^j}\right).$$

$h_N'(1)$ wurde bereits in Satz 4 behandelt.

Nach kurzer Rechnung ergibt sich

$$h_N''(1) = 2 \sum_{k\geq 2} \binom{N}{k} (-1)^{k-1} Q_{k-2} \cdot T(k-2)$$

mit

$$Q_N = \frac{Q(1)}{Q(M^{-N})},$$

wobei

$$Q(z) = \prod_{j\geq 1} \left(1-\frac{z}{M^j}\right)$$

und

$$T(k) = \sum_{1\leq j\leq k} \frac{1}{M^j-1}.$$

Zur asymptotischen Auswertung von $h_N''(1)$ verwenden wir wieder Lemma 3 mit

$$f(z) = -2 \cdot \frac{Q(1)}{Q(M^{-z+2})} \cdot T(z-2),$$

wobei

$$T(z) = \alpha - \sum_{j\geq 1} \frac{1}{M^{z+j}-1}$$

$$\left(\alpha = \sum_{k\geq 1} \frac{1}{M^k-1}\right).$$

Wir bestimmen zunächst die lokale Entwicklung für $u = z-1 \to 0$:

$$\frac{1}{Q(M^{-u+1})} = \frac{1}{1-M^{-u}} \cdot \prod_{k\geq 1} \frac{1}{1-M^{-u-k}}.$$

Ist nun

$$F(u) = \prod_{k\geq 1} \frac{1}{1-f_k(u)} ,$$

so gilt

$$\frac{F'(a)}{F(a)} = \sum_{k\geq 1} \frac{f_k'(a)}{1-f_k(a)}$$

sowie nach kurzer Rechnung

$$\frac{F''(a)}{F(a)} = \left(\frac{F'(a)}{F(a)}\right)^2 + \sum_{k\geq 1} \left(\frac{f_k'(a)}{1-f_k(a)}\right)^2 + \sum_{k\geq 1} \frac{f_k''(a)}{1-f_k(a)} .$$

Wenden wir das mit

$$f_k(u) = M^{-u-k}$$

an, erhalten wir

$$\frac{F'(0)}{F(0)} = - L\cdot\alpha,$$

sowie

$$\frac{F''(0)}{F(0)} = L^2(\alpha^2+\alpha+\beta)$$

mit $L = \log M$, $\alpha = \sum_{k\geq 1} \frac{1}{M^k-1}$, $\beta = \sum_{k\geq 1} \frac{1}{(M^k-1)^2}$.

Demgemäß ergibt sich für $u\to 0$

$$\frac{Q(1)}{Q(M^{-u})} \sim 1 - L\cdot\alpha\cdot u + \frac{1}{2}\cdot L^2\cdot(\alpha^2+\alpha+\beta)u^2.$$

Außerdem haben wir

$$\frac{1}{1-M^{-u}} \sim \frac{1}{L\cdot u}\cdot\left(1+\frac{1}{2}L\cdot u+\frac{1}{12}L^2\cdot u^2\right),$$

sodaß

$$\frac{Q(1)}{Q(M^{-u+1})} = \frac{1}{1-M^{-u}}\cdot\frac{Q(1)}{Q(M^{-u})} \sim \frac{1}{L\cdot u}\left(1+L\left(\frac{1}{2}-\alpha\right)u+\frac{L^2}{2}\left(\alpha^2+\beta+\frac{1}{6}\right)u^2\right) .$$

Nun wenden wir uns der lokalen Entwicklung von $T(u-1)$ für $u\to 0$ zu:

$$T(u-1) = -\frac{1}{M^u-1} + \alpha - \sum_{k\geq 1}\frac{1}{M^{u+k}-1} \sim -\frac{1}{L\cdot u}\left(1-\frac{L}{2}u+\frac{L^2}{12}u^2\right)$$

$$+\alpha - \sum_{k\geq 1}\frac{1}{M^k-1} + L\cdot\sum_{k\geq 1}\frac{M^k}{(M^k-1)^2}\cdot u = -\frac{1}{Lu}+\frac{1}{2}+L\left(-\frac{1}{12}+\alpha+\beta\right)u.$$

Damit erhalten wir insgesamt

$\mathrm{Res}([N;z]f(z);z=1) =$

$$2N\left(\frac{\alpha^2}{2}-\frac{\alpha}{2}-\frac{\beta}{2}-\frac{1}{12}+\frac{\alpha}{L}(1-H_{N-1})+\frac{1}{L^2}\left(1-H_{N-1}+\frac{1}{2}H_{N-1}^2+\frac{1}{2}H_{N-1}^{(2)}\right)\right) .$$

Nun schreiten wir zu den lokalen Entwicklungen für $u = z-\omega_k \to 0$:

$$\frac{Q(1)}{Q(M^{-u-2k\pi i/L})} = \frac{Q(1)}{Q(M^{-u})} \sim 1 - L\cdot\alpha\cdot u,$$

$$\frac{1}{1-M^{-u-2k\pi i/L}} = \frac{1}{1-M^{-u}} \sim \frac{1}{L\cdot u}\left(1 + \frac{Lu}{2}\right),$$

$$T(z-2) = T\left(u-1+\frac{2k\pi i}{L}\right) = T(u-1) \sim -\frac{1}{Lu} + \frac{1}{2}.$$

Also ergibt sich

$$\sum_{k\neq 0} \mathrm{Res}([N;z]f(z);z=\omega_k)$$

$$= \sum_{k\neq 0} \frac{2N^{\omega_k}}{L^2}\left(\log N\cdot\Gamma(-\omega_k)-\Gamma'(-\omega_k)-L\alpha\Gamma(-\omega_k)\right)\cdot$$

$$= 2N(\log_M N-\alpha)\delta^{[D]}(\log_M N) - \frac{2N}{L^2}\sum_{k\neq 0}\Gamma'(-\omega_k)e^{2k\pi i \log_M N}.$$

Durch Kombination dieser Resultate ergibt sich die Behauptung. □

4. EINIGE NUMERISCHE AUSWERTUNGEN

Wir geben noch einmal eine übersichtliche Darstellung der erzielten Ergebnisse und analysieren einige der auftretenden Konstanten genauer.

1. Erwartungswerte

$$\frac{1}{N}E_N - \log_M N \sim$$

$$T : \frac{\gamma}{L} + \frac{1}{2} + \delta^{[T]}(\log_M N)$$

$$P : \frac{\gamma}{L} - \frac{1}{2} + \Delta_M + \delta^{[T]}(\log_M N + \Delta_M)$$

$$D : \frac{\gamma}{L} + \frac{1}{2} - \frac{1}{L} - \alpha + \delta^{[D]}(\log_M N),$$

wobei γ die Eulersche Konstante, $L = \log M$, $\Delta_M = \log_M(M-1)$ und $\alpha = \sum_{k\geq 1}\frac{1}{M^k-1}$ ist.

Vernachlässigt man die periodischen Fluktuationen, so ergeben diese Größen für $M\to\infty$ jeweils den Wert $\frac{1}{2}$. Wir geben im folgenden die Werte für $M = 2,\ldots,5$ an:

	M=2	M=3	M=4	M=5
T	1,33275	1,02540	0,91637	0,85864
P	0,33275	0,65633	0,70885	0,72000
D	-1,71664	-0,56699	-0,22607	-0,06442

2. Varianzen

$$\frac{1}{N} V_N \sim$$

$$T : \frac{1}{12} + \frac{\pi^2}{6L^2} + \sigma^{[T]}(\log_M N)$$

$$P : \frac{1}{12} + \frac{\pi^2}{6L^2} - \frac{2}{L}\mu + \sigma^{[P]}(\log_M N + \Delta_M)$$

$$D : \frac{1}{12} + \frac{\pi^2}{6L^2} + \frac{1}{L^2} - \alpha - \beta + \sigma^{[D]}(\log_M N),$$

wobei

$$\mu = \sum_{n\geq 1} \frac{(-1)^{n-1}}{n(M^n-1)} \text{ und } \beta = \sum_{n\geq 1} \frac{1}{(M^n-1)^2} .$$

Vernachlässigt man die periodischen Fluktuationen, so ergeben diese Größen für $M\to\infty$ jeweils den Wert $\frac{1}{12}$. Wir geben im folgenden die Werte für $M = 2,\ldots,5$ an.

	M=2	M=3	M=4	M=5
T	3,50705	1,44622	0,93926	0,71837
P	1,00000	0,63093	0,50000	0,43068
D	2,84438	1,32532	0,92268	0,73839

Zur numerischen Auswertung von μ bzw. $\alpha+\beta$ (für kleine Werte von M) verwenden wir die folgenden Identitäten, welche in einer nachfolgenden gemeinsam mit J.Schoißengeier verfaßten Note unter Ausnützung von Eigenschaften der Dedekindschen η-Funktion [1] hergeleitet werden:

Bezeichnet

$$f(x) = \sum_{k\geq 1} \frac{(-1)^{k-1}}{k(e^{kx}-1)} ,$$

sodaß

$$\mu = f(\log M),$$

dann gilt:

$$f(x) = \frac{\pi^2}{12x} - \frac{1}{2}\log 2 + \frac{x}{24} - f\left(\frac{2\pi^2}{x}\right) \text{ für alle } x>0.$$

Bezeichne weiters

$$h(x) = \sum_{k\geq 1} \frac{e^{kx}}{(e^{kx}-1)^2} ,$$

sodaß

$$\alpha+\beta = h(\log M),$$

dann gilt:

$$h(x) = \frac{\pi^2}{6x^2} - \frac{1}{2x} + \frac{1}{24} - \frac{2\pi}{x^2}\, h\left(\frac{4\pi^2}{x}\right) \text{ für alle } x>0.$$

LITERATUR:

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

[2] FLAJOLET Ph.und R.SEDGEWICK, Digital Search Trees Revisited, SIAM J.Comput., im Druck.

[3] KIRSCHENHOFER P.und H.PRODINGER, Some Further Results on Digital Search Trees, in: Automata, Languages and Programming (L.Kott ed.), Lecture Notes in Computer Science 226, 177-185, Springer, Berlin 1986.

[4] KNUTH D.E., The Art of Computer Programming Vol.3: Sorting and Searching, Addison-Wesley, Reading Mass.1973.

[5] MEHLHORN K., Data Structures and Algorithms Vol.1, Springer, Berlin 1984.

[6] NÖRLUND N.E., Vorlesungen über Differenzenrechnung, Chelsea, New York 1954.

Peter Kirschenhofer
Helmut Prodinger
Inst.f.Algebra u.Diskrete Mathematik
der Technischen Universität Wien
Wiedner Hauptstr.8-10
1040 WIEN