Nochmals Gleichheitstests

von

Walter Klotz und Wilfried Lex

Abstract:

Procedures testing n-tuples for equal entries are called *equality tests*. The average case complexity is determined explicitly for various equality tests and asymptotic formulas are given.

A, B, C seien Mengen, dabei A geordnet und 0 enthaltend, und es sei

$$A^+ = \{a \in A \mid a > 0\}$$

und

$$C^B = \{f \mid f : B \to C\}$$

sowie $\mathbb{N} = \mathbb{Z}^+$ und

$$\underline{n} = \{x \in \mathbb{N} \mid x \leq n\} \qquad (n \in \mathbb{N}).$$

In Programmen muß gelegentlich geprüft werden, ob eine Liste - etwa der Länge $n(\in \mathbb{N})$ - gleiche Einträge - etwa Zahlen aus $\underline{m}$ mit $m \in \mathbb{N}$ - enthält [3]. Dies läuft auf die Frage nach der Injektivität von t aus $\underline{m}^{\underline{n}}$ hinaus; Algorithmen, die diese Frage beantworten, nennen wir *Gleichheitstests*.

Im folgenden werden einige Ergebnisse zur Komplexität solcher Tests ohne Beweise angegeben, die den Rahmen dieser Zusammenfassung sprengen würden und anderenorts dargestellt werden sollen [2]. – Da die expliziten Formeln für die Durchschnittskomplexität von Gleichheitstests i.a. schon relativ kompliziert sind und sich schwer miteinander vergleichen lassen, werden die Komplexitäten jeweils auch asymptotisch beschrieben.

Für den Trivialfall $m < n$ liefert das Schubfachprinzip sofort die Nichtinjektivität für jedes $t(\in \underline{m}^{\underline{n}})$; dennoch werden die Verfahren auch für $m < n$ untersucht.

I. Vergleichsalgorithmen

Von den insgesamt $\binom{n}{2}!$ Gleichheitstests, die auf direkten Vergleichen beruhen, betrachten wir hier lediglich zwei: $\mathcal{V}$ arbeitet mittels „Vorwärtsvergleichen", indem für $\nu = 1, 2, \cdots$ $\cdots, n-1$ jeweils $t(\nu)$ mit $t(\nu+1), t(\nu+2), \cdots, t(n)$ verglichen und bei $t(\nu) = t(\nu+\rho)$ für $\rho \in \underline{n-\nu}$ abgebrochen wird, während $\mathcal{W}$ „Rückwärtsvergleiche" benutzt, also für $\nu = 2, 3, \cdots, n$ jeweils $t(\nu)$ mit $t(\nu-1), t(\nu-2), \cdots, t(1)$ vergleicht bis zum Abbruch bei $t(\nu) = t(\nu-\rho)$ $(\rho \in \underline{\nu-1})$.

Zur Definition der mittleren Komplexität von $\mathcal{V}$, i.e. Zahl der Vergleiche, sei $v_{m,n} : \underline{m}^{\underline{n}} \to \mathbb{N}_0$ für $m, n \in \mathbb{N}$ sowie für $n > 1$ weiter $v_{m,n}(t)$ die Zahl der notwendigen Vergleiche, um mittels $\mathcal{V}$ das kleinste j aus $\underline{n}$ mit

$$\exists k \in \underline{n} \backslash \underline{j} : t(j) = t(k)$$

zu finden, falls ein solches j existiert, und $v_{m,n}(t) = \binom{n}{2}$ andernfalls; schließlich sei $v_{m,n} = (t \mapsto 0)$. Damit ist

$$V_{m,n} = m^{-n} \sum_{t \in \underline{m}^{\underline{n}}} v_{m,n}(t) \qquad (m, n \in \mathbb{N})$$

die *Durchschnittskomplexität* von $\mathcal{V}$.

In deutlicher Verallgemeinerung von [3], Satz 1, S. 143, erhält man

Satz 1:

Für $m, n \in \mathbb{N}$ gilt

a) $V_{m,n} = m - R_{m,n}$ mit

$$R_{m,n} = \frac{m!}{m^n} \cdot \begin{cases} \left(\frac{m-n}{(m-n)!} + \sum_{\nu=1}^{n} \frac{(m-\nu)^{n-\nu}}{(m-\nu)!} \right) \\ \sum_{\mu=1}^{m-1} \frac{(m-\mu)^{n-\mu}}{(m-\mu)!} \end{cases} \qquad \text{für } m \begin{cases} \geq \\ < \end{cases} n.$$

b) für $R_n = n - V_{n,n}$: $R_n = \frac{n!}{n^n} \sum_{\nu=0}^{n-1} \frac{\nu^\nu}{\nu!}$,

$$1 = R_1 = R_2 > R_3 > \cdots > \frac{1}{e-1} \ ,$$

$$\lim_{n \to \infty} R_n = \frac{1}{e-1} = 0{,}581\ 976 \cdots \ .$$

c) I. $q \in]0,2[\ \wedge \ m \leq n^q \qquad \Rightarrow V_{m,n} - m \in 0(\frac{m}{n})$,

II. $c \in \mathbb{R}^+ \ \wedge \ m - cn^2 \in 0(n) \quad \Rightarrow V_{m,n} - c(1 - e^{-\frac{1}{2c}})n^2 \in 0(n)$,

III. $q \in]2, \infty[\ \wedge \ m \geq n^q \qquad \Rightarrow V_{m,n} - \binom{n}{2} \in 0(\frac{n^4}{m})$.

$w_{m,n}$ und $W_{m,n}$ seien für $\mathcal{W}$ ganz analog zu $v_{m,n}$ und $V_{m,n}$ für $\mathcal{V}$ definiert. Bereits bei $m = n = 4$ differieren die mittleren Komplexitäten von $\mathcal{V}$ und $\mathcal{W}$, und zwar um mehr als 1,46%:

$$W_{4,4} = \tfrac{101}{32} = 3{,}15625 < 3{,}203125 = \tfrac{205}{64} = V_{4,4}.$$

Allgemeineres über diese Abweichung liefert zusammen mit Satz 1 der

Satz 2:

Für $m, n \in \mathbb{N}$ gilt:

a) $W_{m,n} = m^{-n}(\frac{1}{2m}\sum_{\nu=1}^{n}\nu(\nu^2+1)\binom{m}{\nu}\nu!m^{n-\nu} + \binom{m}{n}n!(\binom{n}{2} - \frac{n}{2m}(n^2+1)))$,

für $m < n$ speziell

$$W_{m,n} = m + 1 - \frac{m!}{2m^m}\sum_{\mu=0}^{m}\frac{m^\mu}{\mu!} \quad .$$

b) I. $q \in]0,2[\wedge m \leq n^q \quad \Rightarrow W_{m,n} - m + \frac{1}{2}\sqrt{m\frac{\pi}{2}} \in 0((ln\ m)^2)$,

II. $c \in \mathbb{R}^+ \wedge m - cn^2 \in 0(n) \quad \Rightarrow W_{m,n} - c(1 - e^{-\frac{1}{2c}})n^2 \in 0(n)$,

III. $q \in]2,\infty[\wedge m \geq n^q \quad \Rightarrow W_{m,n} - \binom{n}{2} \in 0(\frac{n^4}{m})$.

Die Beweise der in den Sätzen 1 und 2 vorgestellten Resultate sind länglich; sie benutzen u.a. Rekursionsformeln für Binomialsummen der Form

$$\sum_{k=0}^{n} k^p \binom{m}{k} k! m^{n-k} \qquad (n, p \in \mathbb{N}_0,\ m \in \mathbb{R}).$$

und deren asymptotisches Verhalten [1].

Die Frage nach dem „besten" und „schlechtesten" Vergleichsalgorithmus, d.h. nach dem mit der kleinsten und größten Durchschnittskomplexität, ist offen; ebenso steht eine Klassifizierung aller Vergleichstests noch aus.

II. Lineare Tests

Natürlich gibt es auch Gleichheitstests, bei denen das zu testende n-tupel t nur einmal linear durchwandert werden muß, etwa indem man die Glieder von t als Indizes für ein Boolesches Feld B der Länge m benutzt, dessen Komponenten zunächst alle auf „false" gesetzt sind [3], Def. 38, S. 147; für $\nu = 1, 2, \cdots, n$ wird nun $B[t(\nu)]$ auf „*true*" gesetzt, es sei denn $B[t(\nu)] = \mathit{true}$, in welchem Falle abgebrochen wird. Es handelt sich also i. w. um ein „Hashen ohne Speichern". – Solche Tests mögen *linear* heißen.

Um ihre Komplexität zu beschreiben, sei $l_{m,n} : \underline{m}^{\underline{n}} \to \mathbb{N}$ mit

$$l_{m,n}(t) = \begin{cases} n+1, & \text{falls } t \text{ injektiv ist} \\ min\{\nu \in \underline{n} \mid \exists \mu \in \underline{\nu - 1} : t(\mu) = t(\nu)\}, & \text{sonst} \end{cases}$$

für $t \in \underline{m}^{\underline{n}}$; dann ist

$$L_{m,n} = m^{-n}\sum_{t \in \underline{m}^{\underline{n}}} l_{m,n}(t)$$

die dazugehörige *Durchschnittskomplexität.*

Mit ähnlichen Methoden wie bei den Sätzen 1 und 2 läßt sich zeigen

Satz 3:

Für $m, n \in \mathbb{N}$ gilt:

a) $L_{m,n} = m! \cdot \begin{cases} m^{-m} \sum_{\mu=0}^{m} \frac{m^\mu}{\mu!} \\ m^{-n} \sum_{\nu=0}^{n} \frac{m^{n-\nu}}{(m-\nu)!} \end{cases}$ für $m \left\{ \begin{matrix} \leq \\ > \end{matrix} \right\} n.$

b) I. $m \ ln \ m \leq n^2 \quad \Rightarrow L_{m,n} - \sqrt{m\frac{\pi}{2}} \in 0((ln\ m)^2),$

II. $c \in \mathbb{R}^+ \wedge m - cn^2 \in 0(n) \quad \Rightarrow L_{m,n} - n \int_0^1 e^{-\frac{x^2}{2c}} dx \in 0(1),$

III. $q \in]2, \infty[\wedge m \geq n^q \quad \Rightarrow L_{m,n} - (n+1) \in 0(\frac{n^3}{m}).$

Hierbei ist die Zeit zur Anlage des Feldes und der zeitliche Aufwand zum Aufsuchen der einzelnen Komponenten noch nicht berücksichtigt, so daß die Frage bleibt, welches Verfahren „in praxi" das beste sei, zumal ja auch noch andere Tests existieren.

Literatur

1. W. Klotz: On Certain Binomial Sums. Erscheint in: „Festzeitschrift" in honour of Gerhard Ringel; ed. R. Bodendiek, R. Henn. Physica-Verlag, Heidelberg 1990.

2. W. Klotz, W. Lex: Testing n-tuples for Equal Entries. Eingereicht bei „Discrete Appl. Math." .

3. W. Lex: Einige Bemerkungen zu Gleichheitstests und deren Komplexität. Publication de l'Institut de Recherche Mathématique Avancée, 348/S-17 (1988). Actes du Séminaire Lotharingien de Combinatoire, 17^e Session: 27 - 30 mai 1987, Eremo SS. Pietro e Paolo, Bienno. S. 141 - 148.

Walter Klotz
Institut für Mathematik

Wilfried Lex
Institut für Informatik

der Technischen Universität Clausthal
Erzstr. 1
D-3392 Clausthal-Zellerfeld