Analyse von Hashfunktionen

Borys Gendler

5. Februar 2007

In dieser Arbeit wird die Anzahl der Kollisionen beim Einfügen eines Elements in einer Hashtabelle untersucht. Wir beantworten die Frage, wie sich unterschiedliche Hashfunktion entsprechend diesem Kriterium verhalten. Wir zeigen, dass die sogenannte Gleichverteilung die bestmöglichen Werte hat. Wir beweisen dieses Ergebnis und machen es durch verschiedene Beispiele deutlich.

1 Einleitung, Grundbegriffe des Hashings

Definition 1. *Sei U ein Universum von Elementen und H eine Hashtabelle mit s Stellen. Sei M eine Teilmenge von U. Eine Hashfunktion h weißt jedem Element m der Teilmenge M ein Patz in der Hahtabelle H zu.*

$$h : M \to [0, ..., s-1]$$

Haben zwei Elemente gleiche Positionen zugewiesen bekommen, so spricht man von einer Kollision. Eine gute Hashfunktion soll möglichst wenig Kollisionen verursachen. Es gibt zwei Lösungen des Kollisionsproblems. Man spricht vom offenen Hashing, falls bei der Kollision einfach der nächste freie Platz mit Hilfe von F gesucht wird, und von einem geschlossenen Hashing, falls bei der Kollision verkettete Listen verwendet werden.

Die Funktionen, die wir in der Natur begegnen, verursachen leider eine relativ große Anzahl der Kollisionen. Denkt man an das bekannte Geburtstagsproblem: bei bereits 23 Personen ist es wahrscheinlicher, dass zwei Personen am gleichen Tag Geburtstag haben, als dass alle Geburtstage verschieden sind (selbstverständlich ist Geburtstag an einem Tag gleich einer Kollision). Nimmt man also eine zufällige Funktion, so ist die Wahrscheinlichkeit keiner Kollision, bei dem Versuch 23 Elemente in eine Hashtabelle mit 365 freien Plätzen einzufügen, 0.4927.

Diese grobe und auf keinen Fall vollständige Einführung sollte den Leser an die Grundbegriffe des Hashings erinnern.

2 Gleichverteilung

Sei N die Anzahl der Positionen in der Hashtabelle und k die Anzahl der bereits durch ein Element besetzten Stellen.

Wir betrachten in diesem Abschnitt die sogenannte gleichmäßige Verteilung. Dies bedeutet, dass jedes Element K an einer absolut zufälligen Stelle positioniert wird, d.h. ob eine Stelle bereits besetzt ist nicht damit zusammenhängt, ob die Nachbarn dieser Stelle besetzt sind oder nicht. In diesem Fall sind alle

$$\binom{N}{k}$$

möglichen Konfigurationen absolut gleichwahrscheinlich.

Sei P_r die Wahrscheinlichkeit, dass ein Element genau im r-ten Versuch in die Tabelle mit k besetzten Stellen eingeführt wird. Um P_r zu bestimmen ist folgendes notwendig: die Anzahl der guten Möglichkeiten wird durch die Anzahl aller Kombinationen dividiert. Das Ergebnis lautet:

Hilfssatz 1.

$$P_r = \frac{\binom{N-r}{k-r+1}}{\binom{N}{k}}$$

Beweis. Sei die Menge N sei in zwei disjunkte Mengen A und B geteilt: die Menge der getroffenen A mit r Elementen und die Menge der nicht getroffenen B mit N-r Elementen. In der Menge A sind nun (r-1) besetzte Elemente, und in der Menge B sind dann k-(r-1) besetzte Elemente. □

Definition 2. *Bezeichne C(k, N) die Anzahl der Versuche bis ein Element eingeführt wird, falls in einer Tabelle mit N Elementen k Elementen besetzt sind.*

Satz 1 (über den Erwartungswert)**.**

$$C(k,N) = \frac{N+1}{N-k+1};$$

Beweis.

$$C(k,N) = \sum_{r=1}^{N} rP_r = \underbrace{(N+1)(1 - \underbrace{(P_1 + P_2 + ... + P_N)}_{1})}_{0} + \sum_{r=1}^{N} rP_r$$

$$= N + 1 - (N+1)(\sum_{r=1}^{N} P_r) + \sum_{r=1}^{N} rP_r$$

$$= N + 1 - \sum_{r=1}^{N} (N+1-r)P_r = N + 1 - \sum_{r=1}^{N} (N+1-r)\frac{\binom{N-r}{k-r+1}}{\binom{N}{k}}$$

$$= N + 1 - \sum_{r=1}^{N} (N+1-r)\frac{\binom{N-r}{N-k-1}}{\binom{N}{k}} \overset{(1)}{=} N + 1 - \sum_{r=1}^{N} (N-k)\frac{\binom{N+1-r}{N-k}}{\binom{N}{k}}$$

$$\overset{(2)}{=} N + 1 - (N-k)\frac{\binom{N+1}{N-k+1}}{\binom{N}{k}}$$

$$= N + 1 - (N-k)\frac{(N+1)!(N-k)!k!}{(N-k+1)!k!N!} = N + 1 - (N-k)\frac{N+1}{N-k+1}$$

$$= (N+1)(1 - \frac{N-k}{N-k+1}) = (N+1)(\frac{N-k+1-N+k}{N-k+1}) = \frac{N+1}{N-k+1}$$

□

Erklärung zu (1):

$$\binom{N-r}{N-k-1} = \frac{(N-r)!}{(N-k-1)!(k-r+1)!}\frac{(N-k)(N-r+1)}{(N-k)(N-k+1)} =$$

$$\frac{(N-r+1)!(N-k)}{(N-k)!\underbrace{(k-r+1)}_{N+1-r-N-k}!(N-r+1)} = \binom{N+1-r}{N-k}\frac{N-k}{N-r+1}$$

Erklärung zu (2):

Hilfssatz 2.

$$\sum_{r=1}^{N}\binom{N+1-r}{N-k} = \binom{N}{N-k} + \binom{N-1}{N-k}...\binom{N-k}{N-k} = \binom{N+1}{N-k+1}$$

Beweis. Ansatz: vollständige Induktion nach N und k

Induktionsanfang Sei N = 2, k = 1

$$\binom{2}{1} + \binom{1}{1} = 3 = \frac{3!}{2!1!} = 3$$

Induktionsvoraussetzung:
Gelte die Behauptung

$$\sum_{r=1}^{N}\binom{N+1-r}{N-k} = \binom{N+1}{N-k+1}$$

für alle N und k.

Induktionsschluss:
Zeige, für alle k+1 und N+1 gilt dann

$$\sum_{r=1}^{N+1}\binom{N+2-r}{N-k} = \binom{N+2}{N-k+1}$$

$$\sum_{r=1}^{N+1}\binom{N+2-r}{N-k} = \sum_{r=1}^{N}\binom{N+2-r}{N-k} + \binom{1}{N-k}$$

$$= \binom{N+1}{N-k} + \binom{N}{N-k} + \binom{N-1}{N-k} + ... + \binom{N-2}{N-k}\binom{1}{N-k} \overset{(IV)}{=}$$

$$= \binom{N+1}{N-k} + \binom{N+1}{N-k+1} = \frac{(N+1)!}{(N-k)!(N+1)!} +$$

$$+\frac{(N+1)!}{(N-k+1)!(N)!} = \frac{(N+1)!}{(N-k)!N!(N+1)} + \frac{(N+1)!}{(N-k)!(N-k+1)(N)!}$$

$$= \frac{(N+1)!(N-k+1+k+1)!}{(N-k+1)!(N+1)!} = \frac{(N+2)!}{(N-k+1)!(N+1)!}$$

Analog funktioniert die Induktion für k = k nach N+1, und für N=N nach k+1. Nach diesen drei Schritten ist der Beweis vollbracht. □

3 Optimalitätsanalyse

Sei I eine unendliche Menge der Elemente, von denen eine endliche Teilmenge in die Hashtabelle eingeführt wird. Wir beschreiben eine Hashfunktion h durch die Angabe der Reihenfolge der Stellen, die beim Versuch ein Element in die Tabelle einzufügen nacheinander probiert werden. Wenn h eine Hashfunktion ist, und x ein Elemet, dann bedeutet $h(x) = i_1 i_2 ... i_N$, dass beim Einfügen des Elements i_1 als erste Stelle probiert wird, i_2 als zweite und so weiter bis als N. Stelle i_N kommt.

Es gibt natürlich genau N! Permutationen der Zahlen 1, 2, ...N, also kann die unendliche Menge der Elemente I in die Menge der N! Klassen unterteilt werden. Wir bezeichnen diese Klassen mit $[\pi]$. Jede Klasse $[\pi]$ ist eine Menge von Elementen aus I, mit $h(x) = \pi$. Jede Permutationsklasse $[\pi]$ hat eine gewisse Wahrscheinlichkeit p_π, wobei natürlich die Summe der Wahrscheinlichkeiten aller Permutationsklassen eins ergibt.

Die Frage, die in diesem Abschnitt gestellt wird: welche Verteilung dieser Wahrscheinlichkeiten minimiert $C(k, N)$?

Wie erfolgt das Einfügen eines Elements in die Hashtabelle? Sei $\pi_1, \pi_2, ..., \pi_k$ eine Folge von Permutationen. Sei $\pi_i = j_1, j_2, ..., j_N$. Wenn j_p nicht besetzt ist, und alle $j_1, j_2, ..., j_{p-1}$ besetzt sind, so wird π_i an der Stelle j_p stationiert, und j_p ist damit besetzt.

Beispiel 1. *Seien N=5 und folgende Permutationen gegeben:*

$$\pi_1 = [12345], \pi_2 = [13245], \pi_3 = [54321].$$

Versucht man ein Element aus der Klasse $[\pi_1]$ einzufügen, so wird dieses Element natürlich an der ersten Stelle angesetzt. Damit ist Stelle 1 besetzt. Ist das zweite Element aus der Klasse $[\pi_2]$, so wird es nur an der dritten Stelle landen, da die erste Stelle bereits besetzt ist. Das dritte Element aus der Klasse $[\pi_3]$ besetzt die unbesetzte fünfte Stelle. Damit sind die Stellen 1,3,5 besetzt. Folgende Kombinationen führen zu folgendem Ergebnis:

$\pi_1 \circ \pi_2 \circ \pi_3 = \pi_1 \circ \pi_3 \circ \pi_2 =$ *{1,3,5}*

$\pi_2 \circ \pi_1 \circ \pi_3 = \pi_2 \circ \pi_3 \circ \pi_1 = \{1,2,5\}$
$\pi_3 \circ \pi_1 \circ \pi_2 = \{1,3,5\}$
$\pi_3 \circ \pi_2 \circ \pi_1 = \{1,2,5\}$

Gleichverteilung liegt genau dann vor, wenn $p(\pi_i) = 1/N!$ für alle $i \leq N$ gilt.

Definition 3. *Eine Folge der Permutationen $\sigma = \pi_1\pi_2...\pi_k$ bezeichnen wir als Sequenz.*

Definition 4. *Sei eine Seuqeunz $\pi_1, \pi_2, ..., \pi_k$ gegeben, dann bezeichne $S(\pi_1, \pi_2, ..., \pi_k)$ die Menge der Stellen, die bereits durch Einfügen dieser Sequenz besetzt sind.*

Satz 2. *Die Wahrscheinlichkeit einer Sequenz ist gegeben durch $p_\sigma = p_{\pi_1} \cdot p_{\pi_2} \cdots p_{\pi_k}$.*

(Beweisidee: Jede Permutation hat Ihre eigene Wahrscheinlichkeit, diese Wahrscheinlichkeit hängt nicht von den Wahrscheinlichkeiten der anderen Permutationen ab. Es handelt sich also um eine Kette der unabhängigen Ereignisse).

Satz 3. *Sei nun A eine Untermenge von 1, 2, ..., N, so ist die Wahrscheinlichkeit von A durch folgende Summe für alle σ mit $S(\sigma) = A$ gegeben:*

$$p_A = \sum_{\sigma} p_\sigma$$

Hat eine Menge A die Kardinalität k, so gilt:

$$\sum_{|\sigma|=k} p_\sigma = \sum_{|A|=k} p_A = 1$$

Definition 5. *Bezeichne α eine beliebige Menge unterschiedlicher Symbole zwischen 1 und N. Die Wahrscheinlichkeit von α wird wie folgt bestimmt:*
1) $|\alpha| = N$, so ist $p_\alpha = p_\pi$, π ist die entsprechende Permutation.
2) Wenn $|\alpha| < N$ so ist die Wahrscheinlichkeit wie folgt definiert

$$p_\alpha = \sum_{i \notin \alpha} p_{\alpha i}$$

.

Beispiel 2. *Sei N=4, k=2*
Seien folgende Permutationen mit ihren Wahrscheinlichkeiten gegeben:
p[1234] = p[2143] = 0.2
p[3124] = p[3214] = p[4123] = p[4213] = 0.15
Wie bestimmt man nun p_1?
$p_1 = p_{12} + p_{13} + p_{14} = p_{123} + p_{124} + p_{132} + p_{134} + p_{142} + p_{143} = p_{1234} + \underbrace{p_{1243} + p_{1324} + p_{1342} + p_{1423} + p_{1432}}_{0} = 0.2$

Man kann sagen, dass p_1 die Summe der Wahrscheinlichkeiten aller Permutationen ist, die mit 1 anfangen. Analog ist p_{12} die Summer der Wahrscheinlichkeiten aller Permutationen, die mit [1 2] anfangen und so weiter.

Satz 4.

$$C(k,N) = 1 + \sum_{i=1}^{k} \sum_{A \subseteq 1,2,\ldots,N,\, |A|=k} p_A \sum_{|\alpha|=i,\, \alpha<A} p_\alpha$$

(Die Beweisidee folgt im Kapitel 3.1)

Die 1 in der Formel steht für das Einfügen, wenn ein freies Element gefunden wurde. Für $i = 1$ bedeuten die Summen genau die Wahrscheinlichkeit, dass das erstes Element besetzt ist, für $i = 2$ die Wahrscheinlichkeit, dass die ersten zwei Elementen besetzt sind, und so weiter.

Dabei steht $\alpha < A$ dafür, dass α nur aus den Symbolen besteht, die auch in A vorkommen.

Um das Ergebnis aus Satz 4 zu verdeutlichen folgt ein Beispiel:

Beispiel 3. *Sei N=4, k=2*

Seien folgende Permutationen mit ihren Wahrscheinlichkeiten gegeben:

p[1234] = p[2143] = 0.2

p[3124] = p[3214] = p[4123] = p[4213] = 0.15

Eine Teilmenge kann als Ergebnis einer Folge von Permutationen entstehen, man muss also nur die Wahrscheinlichkeiten dieser Folgen addieren:

{1,2} = [1234] [2143] = [2143] [1234] = [1234] [1234] = [2143] [2143]

$$p\{1,2\} = \underbrace{p[1234]p[2143]}_{0.2*0.2} + \underbrace{p[2143]p[1234]}_{0.2*0.2} + \underbrace{p[1234]p[1234]}_{0.2*0.2} + \underbrace{p[2143]p[2143]}_{0.2*0.2}$$

{1,3} = [1234] [3124] = [3124] [1234] = [1234] [3214] = [3214] [1234] = [3124] [3124] = [3214] [3124]

$$p\{1,3\} = \underbrace{p[1234]p[3124]}_{0.2*0.15} + \underbrace{p[3124]p[1234]}_{0.15*0.2} + \underbrace{p[1234]p[3214]}_{0.2*0.15} +$$
$$\underbrace{p[3214]p[1234]}_{0.15*0.2} + \underbrace{p[3124]p[3124]}_{0.15*0.15} + \underbrace{p[3214]p[3124]}_{0.15*0.15} = 0.165$$

{1,4} = [1234] [4123] = [4123] [1234] = [1234] [4213] = [4213] [1234] = [4123] [4123] = [4213] [4123]

$$p\{1,3\} = \underbrace{p[1234]p[4123]}_{0.2*0.15} + \underbrace{p[4123]p[1234]}_{0.15*0.2} + \underbrace{p[1234]p[4213]}_{0.2*0.15} +$$
$$\underbrace{p[4213]p[1234]}_{0.15*0.2} + \underbrace{p[4123]p[4123]}_{0.15*0.15} + \underbrace{p[4213]p[4123]}_{0.15*0.15} = 0.165$$

{2,3} = [2143] [3124] = [3124] [2143] = [2143] [3214] = [3214] [2143] = [3214] [3214] = [3124] [3214]

$$p\{1,3\} = \underbrace{p[2143]p[3124]}_{0.2*0.15} + \underbrace{p[3124]p[2143]}_{0.15*0.2} + \underbrace{p[2143]p[3214]}_{0.2*0.15} +$$
$$\underbrace{p[3214]p[2143]}_{0.15*0.2} + \underbrace{p[3214]p[3214]}_{0.15*0.15} + \underbrace{p[3124]p[3214]}_{0.15*0.15} = 0.165$$

{2,4} = [2143] [4123] = [4123] [2143] = [2143] [4213] = [4213] [2143] = [4213] [4213] = [4123] [4213]

$$p\{1,4\} = \underbrace{p[2143]p[4123]}_{0.2*0.15} + \underbrace{p[4123]p[2143]}_{0.15*0.2} + \underbrace{p[2143]p[4213]}_{0.2*0.15} +$$
$$\underbrace{p[4213]p[2143]}_{0.15*0.2} + \underbrace{p[4213]p[4213]}_{0.15*0.15} + \underbrace{p[4123]p[4213]}_{0.15*0.15} = 0.165$$

Für {3,4} erfolgt die Berechnung analog.

Damit bekommen wir folgende Werte:

p{1,2} = 0.16

p{3,4}= 0.18

p{1,3} = p{1,4} = p{2,3} = p{2, 4} = 0.165

Nun bestimmen wir die p_αs:

$p_1 = p_{12} + p_{13} + p_{14} = p_{123} + p_{124} + p_{134} + p_{132} + p_{142} + p_{143} = p_{1234} + p_{1243} + p_{1342} + p_{1324} + p_{1423} + p_{1432} = 0.2 + 5 * 0 = 0.2$

Analog bestimme p_i für i = 2, 3, 4.

$p_{12} = p_{123} + p_{124} = p_{1234} + p_{1243} = 0.2$

Analog bestimmt man alle p_αs.

Damit bekommen wir folgende Werte:

p{1,2} = 0.16

p{3,4}= 0.18

p{1,3} = p{1,4} = p{2,3} = p{2, 4} = 0.165

Jetzt sind alle Hilfskomponenten bestimmt. Wir können nun die Formel für C(k, N) anwenden und bekommen:

$$C(k,N) = C(2,4) = 1 + p(1,2)(p_1 + p_2 + p_{12} + p_{21}) + p(1,3)(p_1 + p_3 + p_{13} + p_{31}) +$$
$$+ p(1,4)(p_1 + p_4 + p_{14} + p_{41}) + p(2,3)(p_2 + p_3 + p_{23} + p_{32}) +$$
$$+ p(2,4)(p_2 + p_4 + p_{24} + p_{42}) + p(3,4)(p_3 + p_4 + p_{34} + p_{43}) = 1.665$$

Wie kann man diese Formel verbal beschreiben?

P(i,j) sind die Wahrscheinlichkeiten, dass genau das i. und j. Element besetzt sind. Da alle Kombinationen von i und j genommen werden, sind alle zweielementigen Mengen geprüft. Nun sollen aber besetze Elemente auch von der nächsten Permutation getroffen werden. Die W'keit, dass beim ersten Versuch das i. Element getroffen wird ist durch p_i gegeben, die W'keit, dass beim ersten Versuch das j. Element getroffen wird ist durch p_j gegeben, die W'keit, dass beim ersten und zweiten Versuch das i. und j. Element getroffen werden ist durch p_{ij} gegeben, und letztlich die W'keit, dass beim ersten und zweiten Versuch das j. und i. Element getroffen werden ist durch p_{ji} gegeben. Diese vier Kombinationen beschreiben alle Kombinationen der besetzten Stellen, danach ist die nächste Stelle auf jeden Fall frei!

3.1 Gleichverteilung durch die Optimalitätsanalyse

Jetzt betrachten wir den Fall, dass eine Gleichverteilung vorliegt, also jede Permutation die Wahrscheinlichkeit 1/N! hat. Wir beweisen für diese Verteilung die Aussage aus Satz 4.

Satz 5. *Hat jede Permutation die Wahrscheinlichkeit 1/N!, so gilt:*

$$C(k,N) = 1 + \sum_{i=1}^{k} \sum_{A \subseteq 1,2,\ldots,N,\, |A|=k} p_A \sum_{|\alpha|=i,\, \alpha<A} p_\alpha$$

Beweis. Es gilt für jede Teilmenge A:

$$p_A = \frac{|A|!(N-|A|)!}{N!} = \frac{1}{\binom{N}{|A|}} \tag{1}$$

Wie kommt man zu diesem Wert? Wir können unsere |A| Elemente in beliebiger Reihenfolge platzieren, bekommen also |A|! Möglichkeiten. Analog existieren (N-|A|)! Möglichkeiten, die Elemente, die nicht zu A gehören zu verteilen. Jede Kombination kann mit jeder anderen kombiniert werden, also bekommen wir ein Produkt.

Ähnlich aber etwas anders sieht es mit der Wahrscheinlichkeit p_α aus. Nach der Definition ist p_α die Summe der Wahrscheinlichkeiten aller Permutationen von N Elementen, die mit α anfangen. Es kommt also jetzt auf die Reihenfolge an, es können nur die $(N - |\alpha|)$ Elemente unterschiedliche Stellen besetzen, die ersten $|\alpha|$ Elemente müssen fest bleiben. Wendet man wieder den Ansatz mit guten und schlechten Elementen an, so bekommen wir:

$$p_\alpha = \frac{(N-|\alpha|)!}{N!} \tag{2}$$

Bevor wir die endgültige Antwort geben können, bleibt noch zu klären, wie viele Summanden die Komponenten liefern, also wie oft muss p_α berücksichtigt werden (die Wahrscheinlichkeiten waren gleich für alle A und alle α).

Wie bereits im Beispiel gezeigt, muss bei der Berechnung von p_α zu einer Menge A die Reihenfolge berücksichtigt werden. Es handelt sich also um das Ziehen von i Elementen aus k ohne Zurücklegen mit Berücksichtigung der Reihenfolge. Für jede k-Elementige Teilmenge aus N Elementen haben wir also folgende Anzahl der dazugehörigen αs, mit jeweils in (3) bestimmten Wahrscheinlichkeiten:

$$\sum_{i=1}^{k} \frac{k!}{(k-i)!}$$

Das es insgesamt $\binom{N}{|A|}$ solche Teilmengen gibt, und die Wahrscheinlichkeit von denen jeweils in (2) bestimmt ist, folgt:

$$1 + \sum_{i=1}^{k} \sum_{A \subseteq 1,2,\ldots,N,\, |A|=k} p_A \sum_{|\alpha|=i,\, \alpha<A} p_\alpha = 1 + \sum_{i=1}^{k} \frac{k!}{(k-i)!} p_\alpha \binom{N}{k} p_A$$

$$= 1 + \sum_{i=1}^{k} \frac{k!}{(k-i)!\binom{N}{k}} \binom{N}{k} \frac{(N-i)!}{N!} = 1 + \sum_{i=1}^{k} \frac{k}{N} \frac{k-1}{N-1} (\ldots) \frac{k-i+1}{N-i+1}$$

$$\overset{*}{=} 1 + \frac{k}{N-k+1} = \frac{N+1}{N-k+1} = C(k,N)$$

Zu (*): Dieser Schritt kann mit Induktion bewiesen werden.

□

Besschreibe $C_0(k,N) = \frac{N+1}{N-k+1}$ die Gleichverteilung.

Sei h eine andere Hashfunktion, nicht unbedingt gleichverteilt. Eine wichtige Frage ist: Existiert eine Verteilung der Wahrscheinlichkeiten der Permutationen, so dass $C_h(k,N) < C_0(k,N)$. In unserem Beispiel ist $C_h(2,4) = 1.665$ und $C_0(2,4) = 1.666$. Also: Ja!? Aber: die Werte einer Hashfunktion können nur für gewisse k besser als die Werte der Gleichverteilung sein, aber nicht für alle k. Etwas korrekter formuliert gilt:

Wenn $C_h(k,N) < C_0(k,N)$ gilt, dann existiert ein $k' < k$, so dass $C_h(k',N) < C_0(k',N)$.

Tatsächlich gilt in unserem Beispiel für 1: $C_h(1,4) = 1.26$ und $C_0(1,4) = 1.25$

4 k-uniforme Hashfunktionen

Definition 6. *Eine Hashfunktion h wird k-uniform genannt, wenn für alle Mengen A mit $|A| \leq k$ folgende Bedingung erfüllt ist:*

$$\sum_{|\alpha|=|A|,\,\alpha<A} p_\alpha = \frac{|A|!(N-|A|)!}{N!}$$

Wir erinnern uns: Man kann sagen, dass p_1 die Summe der Wahrscheinlichkeiten aller Permutationen ist, die mit 1 anfangen. Analog ist p_{12} die Summer der Wahrscheinlichkeiten aller Permutationen, die mit [1 2] anfangen und so weiter. Eine Funktion ist k-uniform, wenn die Wahrscheinlichkeit einer Menge nicht von den Elementen der Menge selbst abhängt, sondern von |A| und N. Das würde bedeuten, dass für N=4 folgendes gilt:

$$p_{12} = p_{13} = p_{14} = p_{23} = p_{24} = p_{34} = p_{21} = p_{31} = p_{41} = p_{32} = p_{42} = p_{43}$$

Selbstverständlich sind Hashfunktionen k-uniform, falls jede Permutation gleiche Wahrscheinlichkeit hat. Dies ist ganz leicht zu sehen:

$$\sum_{|\alpha|=|A|,\,\alpha<A} p_\alpha = p_\alpha |A|! = \frac{(N-|A|)!}{N!}|A|!$$

Der Umkehrschluss gilt aber nicht: nicht alle k-uniformen Funktionen sind gleichverteilt:

Beispiel 4. *N=3*

$$p\{123\} = p\{231\} = p\{312\} = 1/3;$$

$$p\{321\} = p\{213\} = p\{132\} = 0;$$

Diese Funktion ist k-uniform für alle k=3, denn es gilt für A=1,2,3:

$$\sum_{|\alpha|=3,\,\alpha<A} p_\alpha = p_{123}+p_{132}+p_{213}+p_{231}+p_{312}+p_{321} = \frac{1}{3}+0+\frac{1}{3}+0+\frac{1}{3}+0 = 1 = \frac{3!}{3!}$$

Analog kan man zeigen, dass dies für alle Kombination von /A/, A und k gilt.

Sei nun h eine k-uniforme Hashfukntion, $|A| = j \leq k$ und i eine natürliche Zahl mit $i \leq j$. Dann ist die folgende Summe unabhänig von A selbst und es gilt:

Satz 6.

$$\sum_{|\alpha|=|A|,\,\alpha<A} p_\alpha = \frac{j!(N-i)!}{(j-i)!N!}$$

Beweis.

$$\sum_{|\alpha|=|A|,\,\alpha<A} p_\alpha = \sum_{B\subseteq A,\,|B|=i}\ \sum_{|\alpha|=i,\,\alpha<B} p_\alpha$$

Aus der Definition von k-uniform folgt:

$$\sum_{|\alpha|=i,\,\alpha<B} p_\alpha = \frac{i!(N-i)!}{N!}$$

Da es genau $\binom{j}{i}$ Untermengen von A gibt folgt:

$$\sum_{|\alpha|=|A|,\,\alpha<A} p_\alpha = \binom{j}{i}\frac{i!(N-i)!}{N!} = \frac{j!i!(N-i)!}{N!i!(j-i)!} = \frac{j!(N-i)!}{N!(j-i)!}$$

□

Beispiel 5. *Sei N=4, Jede Permutation habe die Wahrscheinlichkeit* $1/(N!) = 1/24$. *Sei A=1,2,3, i=2. Dann ist:*

$$\sum_{|\alpha|=2,\,\alpha<\{1,2,3\}} p_\alpha = \sum_{B\subseteq A,\,|B|=2}\ \sum_{|\alpha|=2,\,\alpha<B} p_\alpha = \sum_{|\alpha|=i,\,\alpha<\{1,2\}} p_\alpha + \sum_{|\alpha|=i,\,\alpha<\{2,3\}} p_\alpha +$$

$$+ \sum_{|\alpha|=i,\,\alpha<\{1,3\}} p_\alpha = \sum_{|\alpha|=i,\,\alpha<\{1,2\}} p_\alpha + \sum_{|\alpha|=i,\,\alpha<\{1,3\}} p_\alpha + \sum_{|\alpha|=i,\,\alpha<\{2,3\}} p_\alpha$$

$$= p_{12} + p_{21} + p_{13} + p_{31} + p_{23} + p_{32} = p_{1234} + p_{1243} + p_{2134} + p_{2143} + ... = 1/2$$

Verwendet man die ursprüngliche Definition von p_α *so bekommt man wieder* $1/2$*:*

$$\sum_{|\alpha|=2,\,\alpha<\{1,2,3\}} p_\alpha = p_{12} + p_{21} + p_{13} + p_{31} + p_{23} + p_{32} = ... = 1/2$$

Nun folgt am Ende eine sehr interessante Behauptung ohne Beweis:
Ist h eine k-uniforme Funktion und $|A| = k$. Dann ist $p_A = \binom{N}{k}$
(Beweisidee: Induktion)

5 Abschließende Bemerkung

Wir haben das Problem der Kollisionen beim Hashing ausführlich betrachtet und gezeigt, dass für die mittlere Anzahl der Kollisionen $C(k, N)$ folgendes gilt:

$$C(k, N) = \frac{N+1}{N-k+1}$$

Wir haben dieses Ergebnis ausführlich bewiesen. Ein anderes wichtiges Ergebnis, was wir aber nur für die gleichverteilten Funktionen bewiesen haben, lautete:

$$C(k, N) = 1 + \sum_{i=1}^{k} \sum_{A \subseteq 1,2,\ldots,N,\, |A|=k} p_A \sum_{|\alpha|=i,\, \alpha<A} p_\alpha$$

Am Ende haben wir die wichtige Klasse der k-uniformen Hashfunktionen besprochen.

Literatur

[A Note on the Efficiency of Hashing Funktions, J.D. Ullman, JACM 19, 1972, S. 569-575]

[Scatter storage techniques, R. Morris, Comm. ACM 11, 1 (Jan 1968), 38-44]

[The art of Computer Programming, Volume 3, Donald E. Knuth, Second Edition]

[Introduction to Algorithms, T. H. Cormen, C. E. Leiserson, R. L. Rivest, 1989]

[Uniform Hashing is optimal, A. Yao, JACM 32, 1985, S. 687-693]