2. Der Blum-Floyd-Pratt-Rivest-Tarjan Selektions-Algorithmus

Definition 77

Sei $n \in \mathbb{N}$. Der Median (das „mittlere" Element) einer total geordneten Menge von n Elementen ist deren i-kleinstes Element, wobei

$$i = \left\lceil \frac{n}{2} \right\rceil .$$

Bemerkung: Für gerade n wird manchmal auch $i = \left\lfloor \frac{n}{2} \right\rfloor + 1$ benutzt.

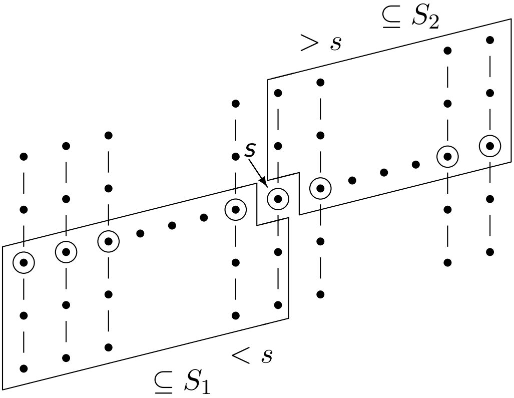
Sei m eine kleine ungerade Zahl (etwa $5 \leq m \leq 21$). Sei $S := \{a_1, \ldots, a_n\}$ eine Menge von n paarweise verschiedenen Elementen. Zur Bestimmung des i-kleinsten Element in S betrachten wir folgenden Algorithmus BFPRT.

Der BFPRT-Selektions-Algorithmus (1/3)

1. Teile S in $\left\lceil \frac{n}{m} \right\rceil$ Blöcke auf, $\left\lfloor \frac{n}{m} \right\rfloor$ davon mit je m Elementen
2. Sortiere jeden dieser Blöcke
3. Sei S' die Menge der $\left\lceil \frac{n}{m} \right\rceil$ Mediane der Blöcke. Bestimme rekursiv den Median s dieser Mediane (also das $\left\lceil \frac{|S'|}{2} \right\rceil$-kleinste Element von S').

Der BFPRT-Selektions-Algorithmus (2/3)

4. Partitioniere $S - \{s\}$ in $S_1 := \{x \in S : x < s\}$, $S_2 := \{x \in S : x > s\}$. Bemerkung: $|S_1|,\ |S_2| \geq \frac{n}{4}$, falls $n \geq 3m - 1$.



Der BFPRT-Selektions-Algorithmus (3/3)

5. Falls $i \leq |S_1|$, bestimme rekursiv das i-kleinste Element in S_1.
   Falls $i = |S_1| + 1$, gib s als Lösung zurück.
   Ansonsten bestimme rekursiv das $(i - |S_1| - 1)$-kleinste Element in S_2.

Sei $T(n)$ die worst-case Anzahl von Vergleichen für $|S| = n$ des Algorithmus BFPRT. Sei C_m die # von Vergleichen, um m Elemente zu sortieren (z.B. $C_5 = 7$, $C_{11} = 26$). Es gilt:

$$T(n) \leq \underbrace{T\left(\left\lceil \frac{n}{m} \right\rceil\right)}_{3.} + \underbrace{T\left(\left\lfloor \frac{3}{4} n \right\rfloor\right)}_{5.} + \underbrace{\left\lceil \frac{n}{m} \right\rceil}_{2.} C_m + \underbrace{\left\lfloor \frac{n}{2} \right\rfloor}_{4.}$$

Satz 78

Der Selektions-Algorithmus BFPRT bestimmt das i-kleinste Element von n Elementen mit $\mathcal{O}(n)$ Vergleichen (und Zeit).

Beweis:

Annahme: $T(n) \le c \cdot n$, wobei $c = c(m)$ konstant ist.

Die Annahme ist ok, falls $T(n) \le$

$$T\left(\left\lceil \frac{n}{m} \right\rceil\right) + T\left(\left\lfloor \frac{3}{4}n \right\rfloor\right) + \left\lceil \frac{n}{m} \right\rceil C_m + \left\lfloor \frac{n}{2} \right\rfloor \le cn; \text{ dies gilt, falls}$$

$$\left\lceil \frac{n}{m} \right\rceil c + \left\lfloor \frac{3}{4}n \right\rfloor c + \left\lceil \frac{n}{m} \right\rceil C_m + \left\lfloor \frac{n}{2} \right\rfloor \le cn \qquad \Big| \cdot \frac{1}{n} \quad \text{(IA)}$$

$$\Leftrightarrow \text{(bis auf } \lceil\,\rceil, \lfloor\,\rfloor \text{)} \ \frac{c}{m} + \frac{3}{4}c + \frac{C_m}{m} + \frac{1}{2} \le c$$

$$\Leftrightarrow -\frac{c}{m} - \frac{3}{4}c + c \ge \frac{C_m}{m} + \frac{1}{2}$$

$$\Leftrightarrow c \ge \frac{\frac{C_m}{m} + \frac{1}{2}}{1 - \frac{3}{4} - \frac{1}{m}}$$

$\square$

Bemerkung: $m = 11 \rightsquigarrow c = c(m) \approx 20$.

Literatur:

Vaughan R. Pratt, Frances F. Yao:
On lower bounds for computing the i-th largest element
Proc. 14th Ann. IEEE SWAT, pp. 70–81 (1973)

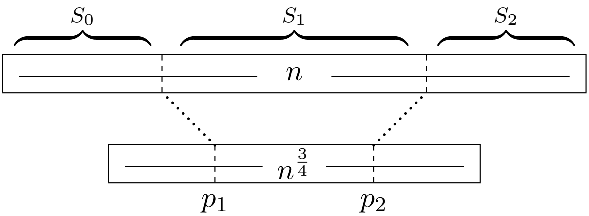
Manuel Blum, Robert W. Floyd, Vaughan R. Pratt, Ron L. Rivest, Robert E. Tarjan:
Time bounds for selection
JCSS **7**, pp. 448–461 (1973)

3. Randomisierter Median-Algorithmus

Problemstellung: Bestimme den Median von n Elementen

1. Wähle $n^{\frac{3}{4}}$ Elemente zufällig und gleichverteilt aus den n Elementen aus.
2. Sortiere diese $n^{\frac{3}{4}}$ Elemente mit einem (Standard-) $n \log n$-Algorithmus.
3. Setze

$$\begin{aligned} p_1 &:= \max\{\tfrac{n^{\frac{3}{4}}}{2} - \sqrt{n}, 1\}\text{-kleinstes Element der } n^{\frac{3}{4}} \text{ Elemente.} \\ p_2 &:= \min\{\tfrac{n^{\frac{3}{4}}}{2} + \sqrt{n}, n^{\frac{3}{4}}\}\text{-kleinstes Element der } n^{\frac{3}{4}} \text{ Elemente.} \end{aligned}$$



4. Partitioniere die n Elemente in

   $$\begin{aligned} S_0 &:= \{\text{Elemente} < p_1\} \\ S_1 &:= \{p_1 \le \text{Elemente} \le p_2\} \\ S_2 &:= \{p_2 < \text{Elemente}\} \end{aligned}$$

5. Falls $|S_0| \geq \lceil \frac{n}{2} \rceil$ oder $|S_2| \geq \lceil \frac{n}{2} \rceil$ oder $|S_1| \geq 4 \cdot n^{\frac{3}{4}}$, dann wiederhole den Algorithmus;
   ansonsten sortiere S_1 und liefere das $(\lceil \frac{n}{2} \rceil - |S_0|)$-kleinste Element davon ab.

Satz 79

Obiger randomisierter Algorithmus bestimmt den Median von n-Elementen mit einer erwarteten Anzahl von $\frac{3}{2}n + o(n)$ Vergleichen.

Beweis:

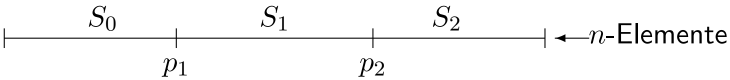
i) Korrektheit: klar.

Beweis (Forts.):

ii) Anzahl der Vergleiche in einer Iteration:

$$\mathcal{O}(n^{\frac{3}{4}} \log n^{\frac{3}{4}}) + \text{Kosten der Partitionierung}$$

Für die Partitionierung ist der naive Ansatz zu ungünstig, stattdessen:



Wähle zuerst jeweils mit Wahrscheinlichkeit $\frac{1}{2}$ aus, ob Element x mit p_1 oder p_2 verglichen wird, mache zweiten Vergleich nur, falls nötig.

Beweis (Forts.):

Die erwartete Anzahl von Vergleichen ist dann

$$\begin{aligned}
&= \frac{n}{2}\left(\frac{|S_0|}{n}\cdot 1 + \frac{|S_1|+|S_2|}{n}\cdot 2\right) + \frac{n}{2}\left(\frac{|S_2|}{n}\cdot 1 + \frac{|S_0|+|S_1|}{n}\cdot 2\right)\\
&= \frac{n}{2}\left(\frac{|S_0|+|S_2|}{n} + 2\frac{\overbrace{|S_0|+|S_1|+|S_2|}^{n}+|S_1|}{n}\right)\\
&= \frac{n}{2}\left(3+\frac{|S_1|}{n}\right) = \frac{3}{2}n + o(n)
\end{aligned}$$

Wir zeigen nun, dass der Algorithmus mit Wahrscheinlichkeit $\geq 1-\mathcal{O}(n^{-\frac{1}{4}})$ nur eine Iteration benötigt (daraus folgt dann, dass insgesamt die Anzahl der Vergleiche $\leq \frac{3}{2}n + o(n)$ ist).

Beweis (Forts.):

Dafür verwenden wir Hilfsmittel aus der Wahrscheinlichkeitstheorie/Stochastik:

- **Bernoulli-Zufallsvariable** (ZV): X, Werte $\in \{0,1\}$ mit

$$X = \begin{cases} 1 \text{ mit WS } p \\ 0 \text{ mit WS } q = 1-p \end{cases}$$

- **Erwartungswert** einer ZV:
  $\mathbb{E}[X] = \sum_{x \in \text{Wertebereich}} x \cdot \Pr[X = x]$
  (X ist diskret, d.h. der Wertebereich von X ist endlich)
- **Markov-Ungleichung**: $\Pr[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$ für X nicht negativ
- **Chebyshev-Ungleichung**: $\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\text{Var}(X)}{t^2}$

Beweis (Forts.):

- **Binomialverteilung**: Seien $X_1, \ldots, X_n$ unabhängige, identisch verteilte Bernoulli-Zufallsvariablen mit $\Pr[X_i = 1] = p$.

$$X := \sum_{i=1}^{n} X_i \ .$$

X ist binomial verteilt, mit Wertebereich $\{0, 1, \ldots, n\}$.

$$\begin{aligned} \Pr[X = k] &= \binom{n}{k} p^k (1-p)^{n-k} \\ \mathbb{E}[X] &= n \cdot p \\ \mathrm{Var}[X] &= n \cdot p \cdot (1-p) = n \cdot p \cdot q \, . \end{aligned}$$

In Zeichen: $X \sim \mathsf{B}(n, p)$

Beweis (Forts.):

Die Auswahl der $n^{\frac{3}{4}}$ Elemente wird wiederholt, falls $|S_0| \geq \frac{n}{2}$. Dies passiert gdw wir höchstens $\frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$ Elemente aus der Hälfte aller Elemente $\leq$ dem Median auswählen.

Wir bestimmen die Wahrscheinlichkeit dafür, dass keine neue Auswahl der $n^{\frac{3}{4}}$ Elemente stattfinden muss.
Setze Bernoulli-Zufallsvariable $X_1, \ldots, X_n$ mit:

$$X_i = \begin{cases} 1, \text{ falls Element } i < \text{Median ausgewählt wird} \\ 0, \text{ sonst} \end{cases}$$

$X := \sum X_i$ ist binomialverteilt mit Parametern n und $\frac{1}{2}n^{-\frac{1}{4}}$, und $\mathbb{E}[X] = \frac{1}{2}n^{\frac{3}{4}}$, $\mathrm{Var}[X] = n \cdot \frac{1}{2}n^{-\frac{1}{4}}(1 - \frac{1}{2}n^{-\frac{1}{4}}) = \frac{1}{2}n^{\frac{3}{4}}(1 - o(1))$.

Beweis (Forts.):

Die Wahrscheinlichkeit hierbei ist

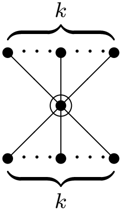
$$\Pr[|S_0| \geq \frac{n}{2}] = \Pr[X \leq \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}] \leq \Pr[|X - \mathbb{E}[X]| \geq \sqrt{n}]$$
$$\leq \frac{\frac{1}{2}n^{\frac{3}{4}}(1 - o(1))}{n} \leq \frac{1}{2}n^{-\frac{1}{4}}(1 - o(1))$$

Die anderen beiden Wahrscheinlichkeitsbedingungen $(\Pr[|S_2| \geq \lceil \frac{n}{2} \rceil]$ und $\Pr[|S_1| \geq 4 \cdot n^{\frac{3}{4}}])$ ergeben analoge Abschätzungen.
Damit: Wiederholung mit WS $\leq \mathcal{O}(n^{-\frac{1}{4}})$. □

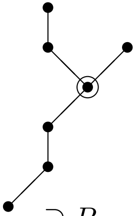
4. Schönhage/Paterson/Pippenger-Median-Algorithmus

Definition 80

Sei $k \in \mathbb{N} \setminus \{0\}$. P_k ist die folgende partielle Ordnung:



$2k + 1$ Elemente

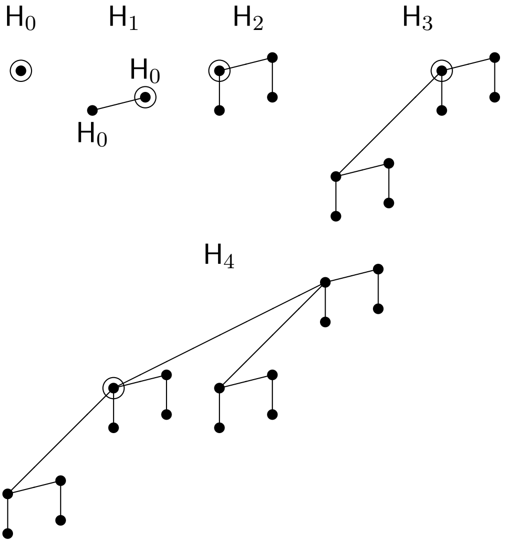


$\supseteq P_3$

Also: Spezielle Binomialbäume mit „Zentrum“.

Definition 81

1. Der Baum H_0 besteht aus einem Knoten, und dieser ist auch das Zentrum.
2. $H_{2h}(h > 0)$ besteht aus zwei H_{2h-1}, deren Zentren durch eine neue Kante verbunden sind. Das Zentrum des H_{2h} ist das kleinere der beiden Zentren der H_{2h-1}.
3. $H_{2h+1}(h \geq 0)$ besteht aus zwei H_{2h}, deren Zentren durch eine neue Kante verbunden sind, sein Zentrum ist das größere dieser beiden Zentren.


H_0
H_1
H_2
H_3
H_0
H_0
H_4

Lemma 82 (Zerlegungslemma)

a) *H_h hat 2^h Knoten, es werden $2^h - 1$ Vergleiche benötigt, um H_h zu konstruieren.*

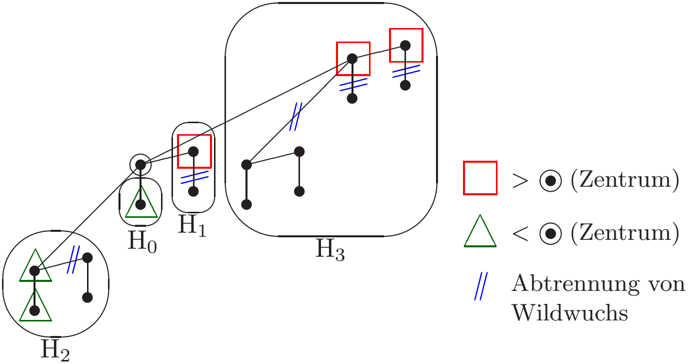
b) *H_{2h} kann zerlegt werden in*
   - *sein Zentrum*
   - *eine Menge $\{H_1, H_3, \ldots, H_{2h-1}\}$ von disjunkten Teilbäumen, deren Zentren alle größer sind als das Zentrum von H_{2h}.*
   - *eine Menge $\{H_0, H_2, H_4, \ldots, H_{2h-2}\}$ von disjunkten Teilbäumen mit Zentren kleiner als das von H_{2h}.*

Lemma 82 (Zerlegungslemma)

c) *H_{2k+1} kann so zerlegt werden, dass die Zusammenhangskomponente des Zentrums genau 2^k Knoten $\geq$ dem Zentrum enthält, indem höchstens $2^{k+1}-1$ Kanten entfernt werden.*
H_{2k} kann so zerlegt werden, dass die Zusammenhangskomponente des Zentrums genau 2^k Knoten enthält, die alle $\leq$ dem Zentrum sind, indem höchstens 2^k-1 Kanten entfernt werden.

d) *Falls $k \leq 2^h-1$, dann kann H_{2h} so zerlegt werden, dass die Zusammenhangskomponente des Zentrums genau $2k+1$ Elemente enthält, von denen k größer und k kleiner als das Zentrum sind ($\Rightarrow P_k$).*
Dazu genügt es, höchstens $3k+2h$ Kanten zu entfernen. Die restlichen Zusammenhangskomponenten sind wieder H_i's.

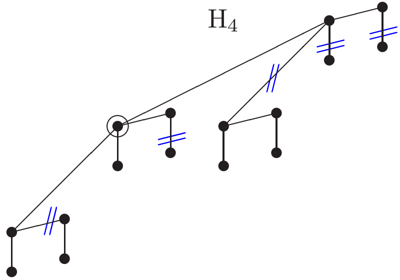
Zerlegungslemma



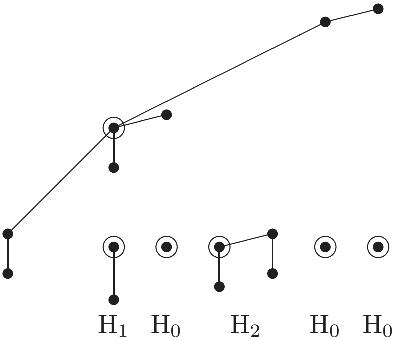
Bemerkung: Bei jedem Konstruktionsschritt wird ein Vergleich durchgeführt, um zu bestimmen, welcher der beiden Teilbäume das kleinere Zentrum hat. Im Algorithmus von Schönhage, Paterson und Pippenger werden aus Teilstücken H_r größere Bäume H_{r+1} zusammengebaut, wodurch schrittweise eine partielle Ordnung auf den Eingabewerten bestimmt wird. Wurde ein Baum H_{2h} hinreichender Größe hergestellt, so wird er durch Zerlegung in einen Baum umgewandelt, der nur noch sein altes Zentrum sowie k darüberliegende und k darunterliegende Elemente enthält, wobei $k \leq 2^h - 1$.

Beispiel 83

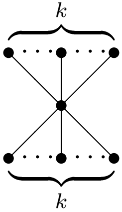
In diesem Beispiel wollen wir H_4 zerlegen und wählen $k = 3$:



Um einen H_4 derart zu zerlegen, müssen wir 5 Kanten aufbrechen. Dabei werden drei H_0, ein H_1 sowie ein H_2 abgespalten.



Übrig bleibt die gewünschte Struktur mit k Knoten über dem Zentrum und k unter dem Zentrum, wodurch eine partielle Ordnung auf $2k+1$ Eingabewerten bestimmt wurde:



$2k+1$ Elemente

Die bei der Zerlegung angefallenen Reststücke werden beim Aufbau weiterer Bäume benutzt. So geht das bereits angesammelte Wissen über die Ordnung der Elemente nicht verloren.

Beweis von a):

Wir beweisen nun die Teile a) bis d) des Zerlegungslemmas.

Lemma 84

H_r hat 2^r Knoten, es werden $2^r - 1$ Vergleiche benötigt, um H_r aufzubauen.

Beweis:

In jedem der r Konstruktionsschritte wird die Anzahl der Knoten verdoppelt. Da wir mit einem Knoten beginnen, hat H_r folglich 2^r Knoten. Die Anzahl der notwendigen Vergleiche C_r unterliegt folgender Rekursionsgleichung ($r \geq 1$):

$$C_r = 1 + 2C_{r-1} \text{ und } C_0 = 0\,.$$

Damit folgt sofort $C_r = 2^r - 1$. $\square$

Beweis von b):

Lemma 85

H_r kann in folgende disjunkte Bereiche unterteilt werden:

- *sein Zentrum,*
- *eine Reihe $H_1, H_3, \ldots, H_{r-1}$ (r gerade) bzw. $\ldots, H_{r-2}$ (r ungerade) von Unterbäumen, deren Zentren über dem von H_r liegen,*
- *eine Reihe $H_0, H_2, \ldots, H_{r-2}$ (r gerade) bzw. $\ldots, H_{r-1}$ (r ungerade) von Unterbäumen, deren Zentren unter dem von H_r liegen.*

Beweis von b):

Beweis:
Durch Induktion über r.
Induktionsanfang: für H_0 gilt die Behauptung.

Induktionsannahme: die Behauptung gelte für H_{r-1}.

Beweis von b):

Beweis:

1. Sei $r = 2h, h > 0$.
   H_{2h} besteht aus zwei H_{2h-1}, wobei das kleinere der beiden alten Zentren das neue Zentrum z bildet. Wende auf den H_{2h-1}, der z enthält, die Induktionsannahme an. Wir können diesen Unterbaum also in z sowie $H_1, H_3, \ldots, H_{2h-3}$ (Zentren über z) und $H_0, H_2, \ldots, H_{2h-2}$ (Zentren unter z) partitionieren. Zusammen mit dem H_{2h-1}, dessen Zentrum über z liegt, ergibt sich die Induktionsbehauptung für H_{2h}.

Beweis von b):

Beweis:

2. Sei $r = 2h + 1$, $h \geq 0$. H_{2h+1} besteht aus zwei H_{2h}, wobei das größere der beiden alten Zentren das neue Zentrum z bildet. Wende auf den H_{2h}, der z enthält, die Induktionsannahme an. Wir können diesen Unterbaum also in z sowie $H_1, H_3, \ldots, H_{2h-1}$ (Zentren über z) und $H_0, H_2, \ldots, H_{2h-2}$ (Zentren unter z) partitionieren. Zusammen mit dem H_{2h}, dessen Zentrum unter z liegt, ergibt sich die Induktionsbehauptung für H_{2h+1}.

$\square$