Übung Algorithm Engineering

Forschungsbericht

Toni Barth und Max Haarbach

3. Oktober 2018

1. Heuristiken

1.1. Heuristik 1: Zufallsdrehungen

Die erste Heuristik führt eine bestimmte Anzahl an Drehungen, die von der Größe der Instanz abhängt, an zufällig ausgewählte Knoten aus. Dieser Vorgang wird wiederum je nach Größe der Instanz mehrfach durchgeführt und am Ende die Sortierung mit dem geringsten Abstand als Ergebnis ausgegeben.

1.2. Heuristik 2: Optimal Leaf Ordering

Die zweite Heuristik nutzt das Verfahren, das Bar-Joseph und weitere für eine möglichst schnelle und optimale Sortierung von hierarchisch geclusterten Datensätzen entwickelt haben [BGJ01]. Dabei wird folgender rekursiver Ansatz verfolgt: Sollen Kosten für einen bestimmten Knoten berechnet werden, setzen sich diese aus den Kosten der beiden Kindknoten und dem Abstand der beiden inneren Blätter dieser beiden Teilbäume. Sofern der Knoten, für denen Kosten berechnet werden sollen, ein Blatt bzw. einen Datensatz darstellt, betragen dessen Kosten 0. Dies ist daher das Rekursionsende. Begonnen wird üblicherweise mit dem Wurzelknoten, da man dadurch am Ende auch die gesamten Abstandskosten berechnet hat.

2. Ziele

Durch die Experimente sollen sowohl

- die Laufzeiten der Heuristiken bei unterschiedlichen Größenordnungen bezüglich der Anzahl der Testobjekte als auch
- die Güte aufgrund der Ähnlichkeiten zu den jeweiligen Originalbildern

ermittelt und verglichen werden.

3. Faktoren

Beim „Leaf-ordering“ sind lediglich 2 Faktoren von Bedeutung:

Zum Einen bestimmt die Größe der Bilder, die im Endeffekt die Anzahl der Testobjekte widerspiegelt, die Laufzeit der Heuristiken. Zum Anderen spielt auch deren Struktur oder Art eine Rolle, die sich allerdings schwer in konkrete Messgrößen oder Werte fassen lassen.

4. Testinstanzen

Gemäß der Faktoren werden auch die Testinstanzen, die durch Grauwert-Bilder realisiert sind, in die entsprechenden Kategorien unterteilt:

- Größen:
  - 10
  - 50
  - 100
  - 500
  - 1000
  - 2500
  - 5000
- Arten:
  - (symmetrische) Testbilder
  - Fotos der realen Welt
  - Farb- bzw. Grau-Übergänge

5. Testergebnisse

Es wurden erst einmal nur Tests für die Instanzen der Größen 10, 50 und 100 durchgeführt, da ab einer Größe von 500 die Laufzeit der zweiten Heuristik stark zunimmt. Dabei wurden insgesamt 3 Durchläufe durchgeführt, in denen jeweils eine Testinstanz mit bestimmter Art und Größe durch die beiden Heuristiken sortiert wurde. Diese Tests wurden auf dem Linux-Server Änubis"der MLU ausgeführt.

5.1. Ergebnis-Qualität

Als Maß für die Qualität der Ergebnisse wird die Summe der Abstände aller benachbarten Blattpaare genutzt, sodass bei 10 Blättern 9 Abstände zu addieren sind. Die Abstände wiederum werden durch den euklidischen Abstand der entsprechenden Spaltenvektoren berechnet. Für die gemischten und sortierten Instanzen sind die Werte im Anhang A zu finden, wobei von letzteren am Ende der Durchschnitt gebildet wird.

Dabei fällt auf, dass die Werte der zweiten Heuristik nicht in allen Fällen gleich sind, wobei der Algorithmus aber immer die beste Lösung finden sollte. Dessen Ursache kann zum Beispiel in Fehlern der Implementation liegen.

5.2. Laufzeiten der Heuristiken

Auch bei den Laufzeiten wurden je Art und Größe der Instanz 3 Messungen durchgeführt, von denen am Ende der Durchschnitt berechnet wird. Die Einheiten der Messungen sind jeweils Sekunden (s). Die Tabellen der Laufzeitenmessungen sind im Anhang B aufgeführt.

6. Wilcoxon Signed-Rank Test

Als Vergleich der beiden Heuristiken wird ein Wilcoxon Signed-Rank Test auf Basis der Durchschnittswerte der Ergebnis-Qualitäten und Laufzeiten der beiden Heuristiken über alle Größen und Arten durchgeführt. Daraus resultieren 18 Datenpunkte pro Heuristik für den Test.

Anhand der Tabelle der kritischen Werte für den Wilcoxon-Test ([Unb]) kann für diese Stichprobengröße und einem Signifikanzniveau von $\alpha = 0,05$ ein kritischer Wert von 40 abgelesen werden. Liegt der aus dem Test ermittelte Wert darunter, befindet er sich im kritischen Bereich und die Nullhypothese, dass die Messwerte ähnlich verteilt sind, kann abgelehnt werden.

Aus Tabelle 1 ist erkennbar, dass ausschließlich positive Ränge vorhanden sind, was auf die zweite Heuristik mit dem Algorithmus der besten Lösung zurückzuführen ist. Daraus resultiert die kleinere Summe der negativen Ränge mit 0 und liegt damit unter dem kritischen Wert.

Größe	Art	Heuristik 1	Heuristik 2	Differenz	Vorz.	Rang	
						pos.	neg.
10	g1	566.623 526 8	426.752 09	139.871 436 8	+	7	
10	g2	308.637 742 5	249.873 800 6	58.763 941 95	+	5	
10	p1	545.985 978	522.444 733 1	23.541 244 98	+	4	
10	p2	396.692 596 4	373.400 786 5	23.291 809 87	+	3	
10	t1	198.422 374 5	192.909 859 2	5.512 515 253	+	1	
10	t2	848.913 536 3	838.263 047 3	10.650 488 97	+	2	
50	g1	2725.438 841	1295.841 883	1429.596 958	+	15	
50	g2	1475.971 813	604.636 533 5	871.335 279 6	+	12	
50	p1	6908.250 429	6587.512 858	320.737 570 4	+	8	
50	p2	6026.672 219	5492.529 012	534.143 207 4	+	9	
50	t1	2935.220 663	2798.969 456	136.251 207 3	+	6	
50	t2	10 971.5055	10 183.3819	788.123 602 8	+	11	
100	g1	4677.816 22	1902.730 879	2775.085 342	+	18	
100	g2	2286.828 587	1072.162 36	1214.666 226	+	14	
100	p1	19 980.725 42	18 371.624 28	1609.101 134	+	16	
100	p2	16 975.622	14 399.115 53	2576.506 477	+	17	
100	t1	10 733.019 92	9955.006 258	778.013 664 1	+	10	
100	t2	19 988.811 25	18 823.779 94	1165.031 308	+	13	
$\sum$						171	0

Tabelle 1: Wilcoxon Signed-Rank Test für Ergebnis-Qualität

Auch bei den Laufzeiten zeigt sich nach Tabelle 2 die kleinere Summe der positiven Ränge mit 21. Dadurch liegt auch diese unter dem kritischen Wert.

Größe	Art	Heuristik 1	Heuristik 2	Differenz	Vorz.	Rang	
						pos.	neg.
10	g1	0.002	0.001	0.001	-	3.5	
10	g2	0.002	0.001	0.001	-	3.5	
10	p1	0.002	0.001	0.001	-	3.5	
10	p2	0.002	0.001	0.001	-	3.5	
10	t1	0.002	0.001	0.001	-	3.5	
10	t2	0.002	0.001	0.001	-	3.5	
50	g1	0.0427	0.3283	0.2856	+		10
50	g2	0.042	0.6557	0.6137	+		12
50	p1	0.043	0.137	0.094	+		8
50	p2	0.043	0.126	0.083	+		7
50	t1	0.0423	0.381	0.3387	+		11
50	t2	0.043	0.1667	0.1237	+		9
100	g1	0.1757	11.404	11.2283	+		15
100	g2	0.1747	14.6813	14.5066	+		17
100	p1	0.1793	5.3723	5.193	+		13
100	p2	0.176	22.9617	22.7857	+		18
100	t1	0.1753	7.7143	7.539	+		14
100	t2	0.1767	12.1773	12.0006	+		16
$\sum$						21	150

Tabelle 2: Wilcoxon Signed-Rank Test für Laufzeiten

Somit agieren die beiden Heuristiken sowohl hinsichtlich der Qualität als auch der Laufzeit sehr unterschiedlich, zumindest in den Größenordnungen von 10 bis 100 Elementen der zu sortierenden Daten.

Des Weiteren ist aus den Werten des Wilcoxon-Test aber auch die eindeutige Tendenz zu einer besseren Laufzeit der ersten Heuristik (21 im Vergleich zu 150) sowie zu einem besseren Ergebnis der zweiten Heuristik (0 im Vergleich zu 171) erkennbar.

A. Ergebnis-Qualität

A.1. Abstandssummen der gemischten Bilder

Art \ Messlauf	#1	#2	#3
g1	834.302 888 687	763.215 338 135	915.014 689 799
g2	523.503 612 176	523.503 612 176	573.212 169 601
p1	715.996 547 763	715.996 547 763	713.029 264 826
p2	428.438 632 521	388.942 148 901	406.337 541 942
t1	248.472 480 733	263.107 387 688	250.725 344 778
t2	974.582 228 968	885.182 192 758	950.296 186 762

Tabelle 3: Abstandssummen der gemischten Instanzen der Größe 10

Art \ Messlauf	#1	#2	#3
g1	3386.103 588 79	2756.287 732 04	3230.964 125 61
g2	2168.001 560 18	2255.887 207 35	2185.994 666 97
p1	7753.696 801	7839.427 282 32	7740.213 454 87
p2	6352.080 464 7	6476.846 136 88	6320.388 645 76
t1	3458.827 822 44	3465.625 281 79	3457.036 366 12
t2	12 478.378 430 9	12 235.026 687 8	12 234.751 462 8

Tabelle 4: Abstandssummen der gemischten Instanzen der Größe 50

Art \ Messlauf	#1	#2	#3
g1	6540.145 814 98	5656.202 094 67	6638.079 235 91
g2	3054.306 395 53	3074.158 430 39	3139.575 751 07
p1	20 765.541 476	20 933.959 189 3	20 762.065 700 1
p2	18 215.889 225 9	17 769.255 600 8	18 113.899 511
t1	12 254.753 481 6	12 363.716 630 1	12 333.007 542 3
t2	20 995.398 447 9	21 076.556 361 5	20 986.002 834 3

Tabelle 5: Abstandssummen der gemischten Instanzen der Größe 100

A.2. Abstandsummen der sortierten Bilder

A.2.1. Heuristik 1

Art \ Messlauf	#1	#2	#3	ø
g1	511.890 729 981	593.989 925 231	593.989 925 231	566.623 526 814 3
g2	378.273 825 497	249.873 800 568	297.765 601 491	308.637 742 518 7
p1	579.379 823 257	522.444 733 052	536.133 377 774	545.985 978 027 7
p2	391.556 113 931	401.822 571 285	396.699 104 033	396.692 596 416 3
t1	207.463 887 593	192.909 859 229	194.893 376 625	198.422 374 482 3
t2	838.263 047 345	869.673 011 642	838.804 549 95	848.913 536 312 3

Tabelle 6: Heuristik 1: Abstandssummen der sortierten Instanzen der Größe 10

Art \ Messlauf	#1	#2	#3	ø
g1	2736.972 577 01	2551.298 143 08	2888.045 802 54	2725.438 840 877
g2	1596.424 495 69	1486.685 810 38	1344.805 133 22	1475.971 813 097
p1	6963.398 126 15	6820.630 855 12	6940.722 304 71	6908.250 428 66
p2	6060.363 256 88	5988.331 808 27	6031.321 592	6026.672 219 05
t1	2834.371 850 44	3058.842 093 43	2912.448 045 9	2935.220 663 257
t2	10 838.680 309 8	10 993.443 405 3	11 082.392 778 2	10 971.505 497 77

Tabelle 7: Heuristik 1: Abstandssummen der sortierten Instanzen der Größe 50

Art \ Messlauf	#1	#2	#3	ø
g1	4835.823 528 44	4756.405 321 19	4441.219 811 18	4677.816 220 27
g2	2214.598 760 15	2293.716 375 33	2352.170 624 42	2286.828 586 63
p1	20 010.819 581 4	19 889.836 909 6	20 041.519 760 8	19 980.725 417 27
p2	17 071.518 985 7	16 979.569 400 1	16 875.777 620 9	16 975.622 002 23
t1	10 585.120 702 7	10 728.523 306 7	10 885.415 757 3	10 733.019 922 23
t2	20 173.403 010 2	20 210.831 500 4	19 582.199 234 3	19 988.811 248 3

Tabelle 8: Heuristik 1: Abstandssummen der sortierten Instanzen der Größe 100

A.2.2. Heuristik 2

Art \ Messlauf	#1	#2	#3	ø
g1	426.752 089 999	426.752 089 999	426.752 089 999	426.752 089 999
g2	249.873 800 568	249.873 800 568	249.873 800 568	249.873 800 568
p1	522.444 733 052	522.444 733 052	522.444 733 052	522.444 733 052
p2	373.400 786 545	373.400 786 545	373.400 786 545	373.400 786 545
t1	192.909 859 229	192.909 859 229	192.909 859 229	192.909 859 229
t2	838.263 047 345	838.263 047 345	838.263 047 345	838.263 047 345

Tabelle 9: Heuristik 2: Abstandssummen der sortierten Instanzen der Größe 10

Art \ Messlauf	#1	#2	#3	ø
g1	1715.866 452 74	1085.829 597 74	1085.829 597 74	1295.841 882 74
g2	604.636 533 543	604.636 533 543	604.636 533 543	604.636 533 543
p1	6587.512 858 29	6587.512 858 29	6587.512 858 29	6587.512 858 29
p2	5492.529 011 7	5492.529 011 7	5492.529 011 7	5492.529 011 7
t1	2798.777 145 1	2799.354 077 59	2798.777 145 1	2798.969 455 93
t2	10 183.381 895	10 183.381 895	10 183.381 895	10 183.381 895

Tabelle 10: Heuristik 2: Abstandssummen der sortierten Instanzen der Größe 50

Art \ Messlauf	#1	#2	#3	ø
g1	1892.391 478	1923.409 680 02	1892.391 478	1902.730 878 673
g2	990.337 428 662	1089.079 858 83	1137.069 793 26	1072.162 360 250 7
p1	18 371.624 283 2	18 371.624 283 2	18 371.624 283 2	18 371.624 283 2
p2	14 399.115 525 6	14 399.115 525 6	14 399.115 525 6	14 399.115 525 6
t1	9955.006 258 08	9955.006 258 08	9955.006 258 08	9955.006 258 08
t2	18 823.779 939 9	18 823.779 939 9	18 823.779 939 9	18 823.779 939 9

Tabelle 11: Heuristik 2: Abstandssummen der sortierten Instanzen der Größe 100

B. Laufzeiten der Heuristiken

B.1. Heuristik 1

Art \ Messlauf	#1	#2	#3	ø
g1	0.002	0.002	0.002	0.002
g2	0.002	0.002	0.002	0.002
p1	0.002	0.002	0.002	0.002
p2	0.002	0.002	0.002	0.002
t1	0.002	0.002	0.002	0.002
t2	0.002	0.002	0.002	0.002

Tabelle 12: Heuristik 1: Laufzeiten bei Instanzen der Größe 10

Art \ Messlauf	#1	#2	#3	ø
g1	0.044	0.042	0.042	0.0427
g2	0.042	0.042	0.042	0.042
p1	0.043	0.042	0.044	0.043
p2	0.043	0.044	0.042	0.043
t1	0.043	0.043	0.042	0.0423
t2	0.043	0.043	0.043	0.043

Tabelle 13: Heuristik 1: Laufzeiten bei Instanzen der Größe 50

Art \ Messlauf	#1	#2	#3	ø
g1	0.176	0.175	0.176	0.1757
g2	0.175	0.174	0.174	0.1747
p1	0.184	0.177	0.177	0.1793
p2	0.175	0.180	0.173	0.176
t1	0.175	0.176	0.175	0.1753
t2	0.176	0.180	0.174	0.1767

Tabelle 14: Heuristik 1: Laufzeiten bei Instanzen der Größe 100

B.2. Heuristik 2

Art \ Messlauf	#1	#2	#3	ø
g1	0.001	0.001	0.001	0.001
g2	0.001	0.001	0.001	0.001
p1	0.001	0.001	0.001	0.001
p2	0.001	0.001	0.001	0.001
t1	0.001	0.001	0.001	0.001
t2	0.001	0.001	0.001	0.001

Tabelle 15: Heuristik 2: Laufzeiten bei Instanzen der Größe 10

Art \ Messlauf	#1	#2	#3	ø
g1	0.327	0.336	0.322	0.3283
g2	0.638	0.668	0.661	0.6557
p1	0.136	0.135	0.140	0.137
p2	0.125	0.127	0.126	0.126
t1	0.381	0.373	0.389	0.381
t2	0.166	0.166	0.167	0.1667

Tabelle 16: Heuristik 2: Laufzeiten bei Instanzen der Größe 50

Art \ Messlauf	#1	#2	#3	ø
g1	10.436	13.532	10.244	11.404
g2	14.771	14.677	14.596	14.6813
p1	5.441	5.334	5.342	5.3723
p2	22.941	22.983	22.961	22.9617
t1	8.215	7.448	7.480	7.7143
t2	12.111	12.275	12.146	12.1773

Tabelle 17: Heuristik 2: Laufzeiten bei Instanzen der Größe 100

Tabellenverzeichnis

Literatur

[BGJ01] Ziv Bar-Joseph, David K Gifford und Tommi S Jaakkola.
„Fast optimal leaf ordering for hierarchical clustering“.
In: *Bioinformatics* 17.suppl_1 (2001), S22–S29.
URL: https://watermark.silverchair.com/17S022.pdf (siehe S. 1).

[Unb] Unbekannt. *Table of critical values for the Wilcoxon test*. University of Sussex.
URL: http://users.sussex.ac.uk/~grahamh/RM1web/WilcoxonTable2005.pdf
(siehe S. 3).