SIGMOD 2010 RWE Review on Paper "Histograms Reloaded: The Merits of Bucket Diversity"

Zhenjie Zhang,

National University of Singapore
zhenjie@comp.nus.edu.sg

1 Experiment Overview

In this report, we will conduct some analysis on the repeatability and workability on the source codes provided by the authors of SIGMOD 2010 paper "Towards Histograms Reloaded: The Merits of Bucket Diversity". This report consists of three parts, including a general summary of comments from primary reviewer in Section 2, an overall repeatability evaluation in Section 3, and some extended experimental results based on a new data set generated by the secondary reviewer in Section 4.

Based on the source codes and running scripts provided by the authors, we have set up some experimental environment on a Red hat Linux Operating system (CentOS 5.0) equipped on IBM x255 server with four Intel Xeon MP 3.0 GHz CPU, 18G DDR memory and six 73.4GB Ultra320 SCSI hard disks. All the programs are compiled with GCC 4.4.3 and each process is handled by a single core at any time.

2 Summary from Primary Reviewer

Generally speaking, the primary reviewer was able to reproduce all of the experiments, based on the program package provided by the author. However, some of the experiments listed in the original paper are not included in the repeatability evaluation, due to the lack of scripts and data, covering Table 1, Table 4 and Table 5. The second reviewer repeated this group of experiments on a new experimental setting and will report the results in Section 3.

On the other hand, the primary reviewer tested the workability by generating some 1-dimensional histograms from TPC-H SF1 data. Some experiments show that the proposed histogram construction method turns out to be very slow when there are a large number of bins with positive frequencies. However, TPC-H data set is famous for its uniformity on the value distributions. To verify this conclusion, the second reviewer has conducted a new experiment on the provided program with a new skewed data set. Details on the data generation and experimental results will be discussed in Section 4.

3 Repeatability Evaluation

Since the authors have provided a nicely organized makefile supporting easy experiments evaluation, result summarization and pdf generation, the second reviewer just runs makefile to repeat the experiments included in the scripts. In the attached file, the results on all the 6 data sets are listed in Section 1.1 and Section 1.3-1.6. All the results are fully satisfactory to the reviewer. The efficiencies of the programs are improved by about 10% on those from primary reviewer, due to the various experimental environment and machine settings.

3 Workability Evaluation

To test the workability of the program package, especially on the existence of large number of non-empty bins, the secondary reviewer generates a skewed synthetic data set as follows. A distribution of Gaussian Mixture Model is created, with 4 different clusters on the 1-dimensional space. Each cluster follows some Gaussian distribution, with a randomly selected center and an identical variance parameter. The sizes of the clusters are carefully controlled so that the biggest cluster is no larger than two times of the smallest cluster. After the creation of the distribution, one million of samples are generated from the distribution, uniformly mapped to an integer domain from 0 to 9,999. Therefore, there are 10,000 bins absorbing the samples, leading to 8,937 non-empty bins. This data set is fed into the program package and the results are listed in Section 1.2 in the attachment, all of which rely on the scripts from the author.

There are several important observations on the experimental results. First, the computation time of the program is fairly slow. It takes about 15 minutes to finish the histogram construction when NoCoeffLimit=3. Second, there is a sharp decrease on the program efficiency when we jumps from q-err=1 to q-err=2, regardless of the other parameters. When q-err is larger than 2, the CPU time is almost a constant even when we expand the tolerance factor q-err. To the reviewer's opinion, this is due to the skewed property of the clustered data, rending small differences on the histograms when the error tolerance is high. The reviewer believes that this needs more investigations from the authors for clearer explanations

4 Conclusion

Generally speaking, the second reviewer is satisfied with the repeatability of the experiments. But the scalability of the method remains in doubt and more research investigations are in need.

Profiles for Single Attributes

Guido Moerkotte

August 12, 2010

1 Profile Summaries

1.1 Profile Summaries for `citeseer`

citeseer, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	21	1423	6	4.386
2	24	1120	5	7.135
2.3	28	911	4	8.049
2.5	33	863	4	9.226
2.7	28	784	3	14.552
3	32	664	3	22.388
3.3	29	597	2	29.387
3.5	32	576	2	35.667
3.7	33	556	2	35.915
4	28	469	2	40.535
4.5	28	408	2	38.898
5	24	328	1	41.786

citeseer, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	15	1373	6	10.298
2	27	1125	5	17.521
2.3	29	901	4	25.196
2.5	25	824	3	31.893
2.7	25	746	3	53.442
3	26	635	3	58.912
3.3	21	546	2	67.094
3.5	21	489	2	87.422
3.7	21	471	2	93.112
4	25	420	2	93.641
4.5	19	314	1	94.665
5	14	220	1	121.44

citeseer, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	15	1373	6	16.884
2	27	1125	5	32.85
2.3	29	901	4	42.755
2.5	24	828	3	56.284
2.7	24	750	3	88.321
3	26	635	3	91.722
3.3	21	546	2	105.503
3.5	20	480	2	138.879
3.7	22	470	2	145.035
4	23	389	2	150.563
4.5	18	307	1	150.794
5	14	220	1	202.761

1.2 Profile Summaries for `data`

data, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	100	3493	3	96.576
2	34	438	0	249.608
2.3	32	391	0	253.29
2.5	32	386	0	253.67
2.7	32	385	0	254.063
3	32	384	0	253.473
3.3	32	384	0	252.299
3.5	32	384	0	251.805
3.7	32	384	0	251.768
4	32	384	0	252.901
4.5	32	384	0	252.109
5	32	384	0	252.276

data, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	104	3515	3	215.519
2	33	424	0	526.679
2.3	32	391	0	562.024
2.5	32	386	0	565.728
2.7	32	385	0	563.274
3	32	384	0	566.071
3.3	32	384	0	566.632
3.5	32	384	0	566.197
3.7	32	384	0	566.208
4	32	384	0	565.73
4.5	32	384	0	563.565
5	32	384	0	566.225

data, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	104	3515	3	337.298
2	33	424	0	825.272
2.3	32	391	0	881.746
2.5	32	386	0	881.56
2.7	32	385	0	889.356
3	32	384	0	890.08
3.3	32	384	0	881.131
3.5	32	384	0	883.646
3.7	32	384	0	889.927
4	32	384	0	885.841
4.5	32	384	0	884.081
5	32	384	0	887.849

1.3 Profile Summaries for `ecb_usdeur`

ecb_usdeur, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	403	8562	33	6.174
2	466	7629	30	6.97
2.3	439	6657	26	6.444
2.5	391	5856	23	5.875
2.7	351	5189	20	5.368
3	290	4237	16	5.002
3.3	240	3520	14	4.806
3.5	209	3076	12	5.188
3.7	186	2696	10	5.734
4	151	2222	9	6.773
4.5	113	1601	6	10.63
5	79	1183	5	15.805

ecb_usdeur, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	402	8558	33	12.59
2	466	7634	30	14.329
2.3	422	6558	25	12.863
2.5	350	5636	22	11.118
2.7	300	4866	19	10.353
3	225	3864	15	9.865
3.3	176	3156	12	10.317
3.5	146	2669	10	11.708
3.7	128	2336	9	15
4	95	1820	7	21.115
4.5	62	1216	5	38.106
5	49	919	4	66.414

ecb_usdeur, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	402	8558	33	19.254
2	466	7634	30	22.059
2.3	422	6558	25	19.993
2.5	350	5636	22	17.732
2.7	300	4866	19	16.807
3	225	3864	15	16.512
3.3	175	3153	12	18.457
3.5	146	2682	10	22.186
3.7	124	2307	9	30.216
4	86	1758	7	37.947
4.5	55	1161	5	74.656
5	42	890	3	110.829

1.4 Profile Summaries for `uniprotAA`

uniprotAA, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	36	1330	4	8.305
2	33	992	3	19.866
2.3	22	906	3	26.198
2.5	29	802	3	28.196
2.7	36	718	2	34.858
3	34	629	2	37.994
3.3	27	573	2	42.309
3.5	27	538	2	42.585
3.7	25	495	2	49.887
4	23	450	1	47.432
4.5	19	400	1	64.839
5	26	373	1	65.016

uniprotAA, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	26	1261	4	23.3
2	28	952	3	43.183
2.3	25	917	3	62.536
2.5	32	791	3	69.151
2.7	33	702	2	83.7
3	29	626	2	86.876
3.3	24	563	2	95.191
3.5	23	521	2	99.726
3.7	20	476	2	103.022
4	19	450	1	96.721
4.5	18	392	1	133.842
5	21	376	1	135.839

uniprotAA, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	26	1266	4	27.765
2	28	956	3	77.628
2.3	24	914	3	106.922
2.5	32	802	3	112.425
2.7	30	693	2	119.205
3	29	631	2	129.842
3.3	24	569	2	156.712
3.5	23	521	2	160.801
3.7	20	476	2	163.353
4	18	441	1	155.115
4.5	18	392	1	207.745
5	22	378	1	211.825

1.5 Profile Summaries for `uniprotMW`

uniprotMW, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	286	50638	6	135
2	688	39273	4	123.55
2.3	912	34613	4	162.248
2.5	1152	30828	3	332.918
2.7	1024	25526	3	809.108
3	813	19668	2	1217.7
3.3	829	17993	2	1457.47
3.5	802	17304	2	1516.6
3.7	815	16403	2	1623.47
4	775	14505	2	1730.3
4.5	786	12854	1	1836.12
5	756	11100	1	2073.72

uniprotMW, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	272	50510	6	262.757
2	692	39278	4	247.095
2.3	805	33984	4	319.695
2.5	976	29592	3	669.801
2.7	819	23819	3	1669.93
3	689	19007	2	2511.64
3.3	650	16908	2	3048.85
3.5	625	15883	2	3195.93
3.7	597	14656	2	3572.12
4	637	13476	1	3869.59
4.5	578	10673	1	4478.2
5	466	8118	1	5348.95

uniprotMW, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	275	50510	6	400.925
2	693	39283	4	389.304
2.3	795	33969	4	507.311
2.5	978	29616	3	1041.38
2.7	816	23814	3	2586.29
3	688	19026	2	3863.41
3.3	647	16867	2	4769.74
3.5	614	15832	2	5020.93
3.7	581	14503	2	5659.04
4	602	13095	1	6136.13
4.5	535	10289	1	7203.6
5	436	7863	1	9049.42

1.6 Profile Summaries for `wtrslp`

wtrslp, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	370	5838	26	1.785
2	352	5231	23	1.796
2.3	152	2159	10	2.535
2.5	22	330	1	31.744
2.7	20	294	1	35.525
3	17	244	1	35.715
3.3	44	591	3	11.837
3.5	14	193	1	43.492
3.7	35	462	2	19.398
4	29	388	2	19.067
4.5	22	301	1	26.041
5	18	241	1	46.632

wtrslp, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	370	5838	26	4.103
2	227	4039	18	3.792
2.3	129	2024	9	9.283
2.5	13	255	1	86.537
2.7	11	220	1	85.149
3	10	194	1	76.176
3.3	41	574	3	32.182
3.5	8	149	1	113.694
3.7	32	459	2	44.299
4	27	367	2	48.667
4.5	21	288	1	68.682
5	18	241	1	104.911

wtrslp, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	370	5838	26	6.587
2	227	4039	18	6.398
2.3	129	2028	9	16.111
2.5	12	257	1	129.649
2.7	11	224	1	140.177
3	10	194	1	156.56
3.3	41	574	3	53.028
3.5	8	161	1	195.636
3.7	32	459	2	81.11
4	27	367	2	83.491
4.5	21	288	1	112.328
5	18	241	1	167.04

1.7 Profile Summaries for `wtrtmp`

wtrtmp, NoCoeffLimit=1				
q-err	#buckets	size	bits	cpu
1.7	457	6949	25	2.579
2	398	5639	20	2.755
2.3	246	3474	12	4.802
2.5	143	2021	7	10.278
2.7	99	1359	5	29.114
3	67	900	3	43.545
3.3	46	615	2	45.12
3.5	36	502	2	70.889
3.7	34	455	2	76.042
4	27	372	1	79.2
4.5	21	290	1	98.86
5	19	264	1	96.653

wtrtmp, NoCoeffLimit=2				
q-err	#buckets	size	bits	cpu
1.7	457	6949	25	5.688
2	357	5210	19	6.409
2.3	136	2390	9	23.029
2.5	87	1515	5	45.999
2.7	69	1180	4	73.427
3	51	834	3	121.62
3.3	40	609	2	134.92
3.5	34	490	2	187.001
3.7	31	456	2	212.307
4	26	381	1	218.151
4.5	21	308	1	244.892
5	18	264	1	245.669

wtrtmp, NoCoeffLimit=3				
q-err	#buckets	size	bits	cpu
1.7	457	6949	25	9.125
2	357	5210	19	10.902
2.3	136	2390	9	43.188
2.5	87	1515	5	86.631
2.7	69	1180	4	122.211
3	51	834	3	207.009
3.3	40	609	2	230.451
3.5	34	490	2	303.272
3.7	31	456	2	350.497
4	26	381	1	355.698
4.5	21	308	1	401.296
5	18	272	1	394.616

2 Excerpts

2.1 Excerpts From Homogeneous Profiles

attr	T—-	TM—	T-Q–	TMQ–	T-QT-	TMQT-
citeseerciteseer data ecb_usdeurecb_usdeur uniprotAAuniprotAA uniprotMWuniprotMW wtrslpwtrslp wtrtmpwtrtmp	6981	9197	1719	2361	10311	12575

2.2 Excerpts From Heterogeneous Profiles, NoCoeffLimit=1

Heterogeneous Profile, NoCoeffLimit = 1						
attr	nodistval	size	bits	qerr	size	bits
citeseer						
data	9430	438	0			
ecb_usdeur						
uniprotAA						
uniprotMW						
wtrslp						
wtrtmp						

2.3 Excerpts From Heterogeneous Profiles, NoCoeffLimit=2

Heterogeneous Profile, NoCoeffLimit = 2						
attr	nodistval	size	bits	qerr	size	bits
citeseer						
data	9430	424	0			
ecb_usdeur						
uniprotAA						
uniprotMW						
wtrslp						
wtrtmp						

2.4 Excerpts From Heterogeneous Profiles, NoCoeffLimit=3

Heterogeneous Profile, NoCoeffLimit = 3						
attr	nodistval	size	bits	qerr	size	bits
citeseer						
data	9430	424	0			
ecb_usdeur						
uniprotAA						
uniprotMW						
wtrslp						
wtrtmp						