International Journal Data Science and Analytics manuscript No.
(will be inserted by the editor)

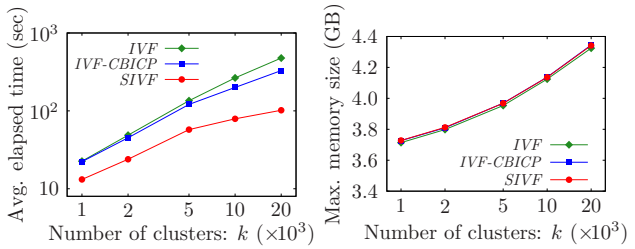
Supplementary Material: Structured Inverted-File k-Means Clustering for High-Dimensional Sparse Data

Kazuo Aoyama · Kazumi Saito

Abstract This provides the experimental results when the algorithms described in the paper were executed by 50-thread processing for NYT (data set) described in Section 5.1. Section 9 shows the results of the three algorithms of *SIVF*, *IVF-CBICP*, and *IVF*. Sections 10 and 11 show the results of the four algorithms of *SIVF*, *Ding*$^+$, *Drake*$^+$, and *Lloyd-ICP*. In Fig. 12 in Section 10, the five data sets with different sizes were used while the data set with 1 285 944 objects (texts), which had the maximum size, was used in the other results. All the results in NYT were similar to those in PubMed.

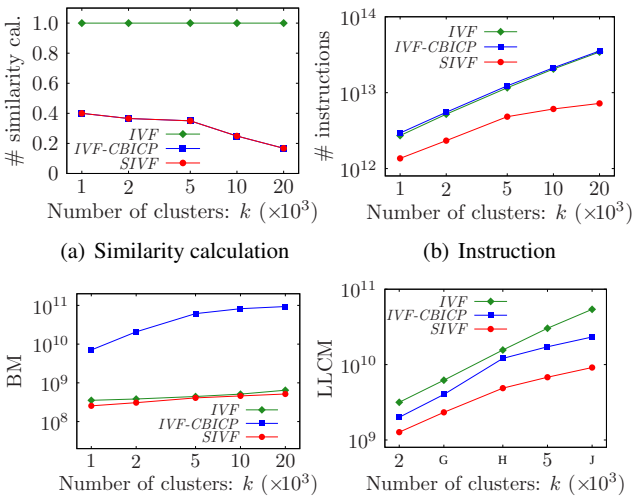
9 Properties of SIVF in NYT

Figures 9 and 10 correspond to Figs. 2 and 3 in Section 6.



(a) Avg. elapsed time (b) Max. memory size

Fig. 9 Performance of three algorithms that were executed by 50-thread processing with given k in NYT. (a) Average elapsed time per iteration is plotted along k with log-log scale and (b) Occupied maximum memory size through iterations along k with log-linear scale.



(a) Similarity calculation (b) Instruction

(c) Branch misprediction (d) LLC miss

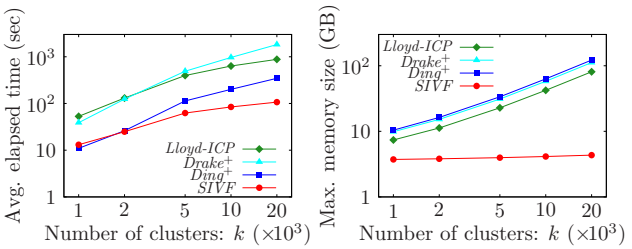
Fig. 10 (a) Number of similarity calculations normalized by $(N \times k)$ and characteristics of performance degradation factors: (b) the number of instructions (Inst), (c) branch mispredictions (BM), and (d) last-level cache misses (LLCM) when the three algorithms were executed by 50-thread processing for NYT. (a) is plotted with log-linear scale and (b), (c), and (d) with log-log scale.

K. Aoyama
NTT Communication Science Laboratories, 2-4, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0237 Japan.
Tel.: +81-774-93-5270
E-mail: kazuo.aoyama.rd@hco.ntt.co.jp

K. Saito
Kanagawa University, 2946, Tsuchiya, Hiratuka-shi, Kanagawa 259-1293 Japan.

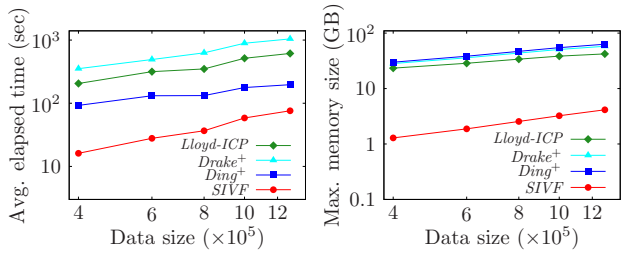
10 Performance Comparison in NYT

Figures 11 and 12 correspond to Figs. 4 and 5 in Section 7.1.



(a) Avg. elapsed time (b) Max. memory size

Fig. 11 Performance of four algorithms executed by 50-thread processing with given k in NYT. (a) Average elapsed time per iteration and (b) Occupied maximum memory size through iterations were plotted along k with log-log scale.



(a) Avg. elapsed time (b) Max. memory size

Fig. 12 Performance of four algorithms executed by 50-thread processing for NYT with different sizes ($4\times10^5, 6\times10^5, 8\times10^5, 1\times10^6, 1\,285\,944$) when $k = 10\,000$. (a) Average elapsed time per iteration and (b) Occupied maximum memory size through iterations are plotted along data size with log-log scale.

11 Performance Analysis in NYT

Figures 13 and 14 correspond to Figs. 6 and 7 in Section 7.2, respectively. Figure 15 corresponds to Fig. 8.

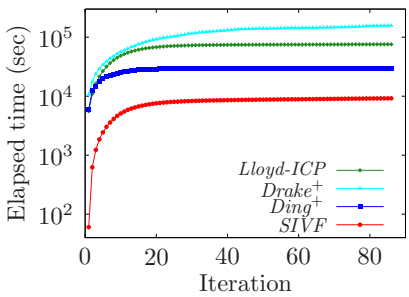


Fig. 13 Elapsed time that four algorithms with $k = 20\,000$ required until convergence when they were applied to NYT. Elapsed time is plotted along iteration with linear-log scale.

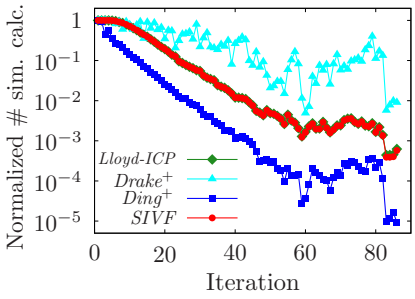
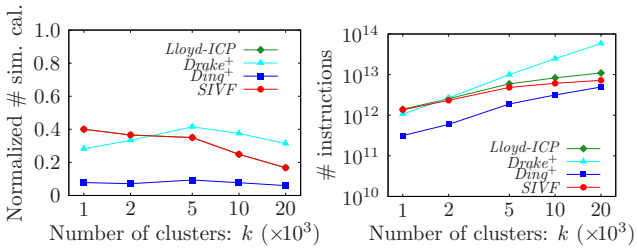
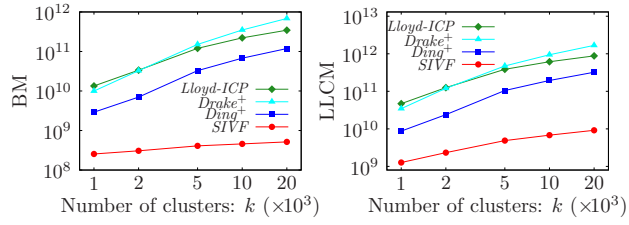


Fig. 14 Number of similarity calculations normalized by ($N\times k$), where N indicates the data size of 1 285 944, when four algorithms with $k = 20\,000$ were executed by 50-thread processing for NYT. Normalized number is plotted along iteration with linear-log scale.



(a) Similarity calculation (b) Instruction



(c) Branch misprediction (d) LLC miss

Fig. 15 (a) Number of similarity calculations normalized by ($N\times k$) and characteristics of performance degradation factors: (b) the number of instructions (Inst), (c) branch mispredictions (BM), and (d) last-level cache misses (LLCM) when the three algorithms were executed by 50-thread processing for NYT. (a) is plotted with log-linear scale and (b), (c), and (d) with log-log scale.