Supplementary Materials for Paper: "Automatic Taxonomy Construction from Keywords via Scalable Bayesian Rose Trees"

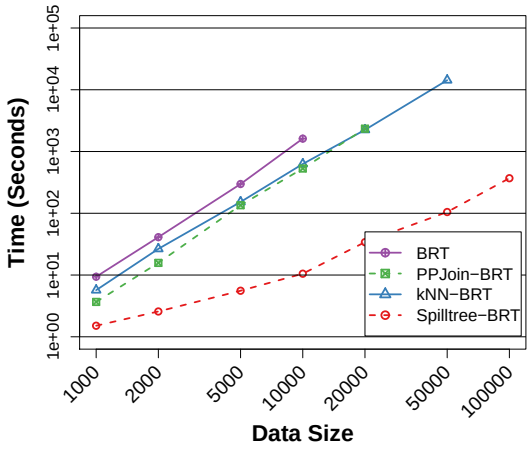
Yangqiu Song, Shixia Liu, Xueqing Liu, and Haixun Wang

September 30, 2014

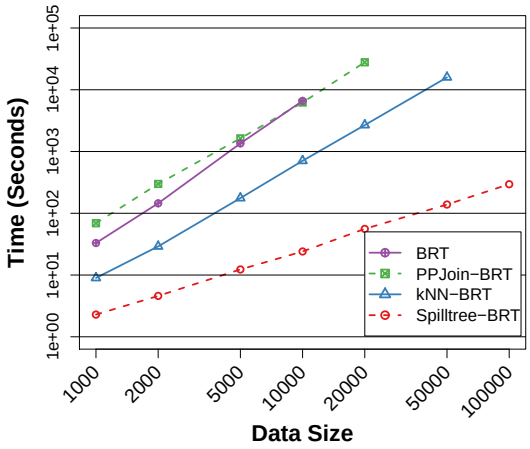
Additional Experiments

To further analyze the performance of each algorithm in detail, we divided the execution into two parts: preprocessing and build. The detailed time costs for different algorithms is shown in Figs. 1(a), 1(b), 1(c), 1(d) and Figs. 2(a), 2(b), 2(c), 2(d) for DCM and vMF distribution modeling respectively. For BRT, during the preprocessing, we computed the likelihood values between data samples and then sort them. For the nearest neighbor based methods, the preprocessing mainly focused on finding the nearest neighbors and computing the likelihood values. The building part of each algorithm gradually merges clusters. For different algorithms, merging clusters differs from searching for the candidate pair sets of different sizes. Moreover, after a new cluster is generated, a different number of likelihood values will be computed. It can be concluded from the results that the time costs of kNN-BRT and Spilltree-BRT in these two parts is consistently lower than BRT. Specifically, the preprocessing of PPJoin-BRT is faster than kNN-BRT. However, the build is much slower. This again demonstrates our explanation that PPJoin-BRT will re-run the process of finding the nearest neighbors.

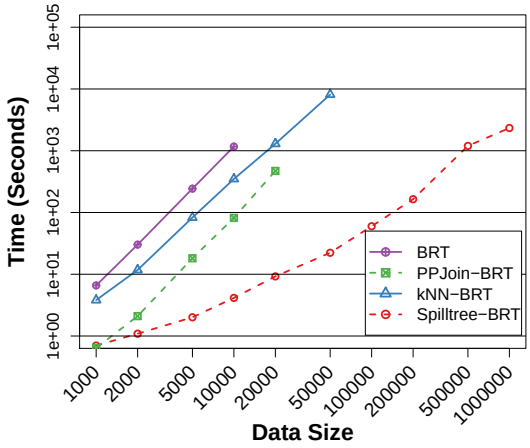
Moreover, we evaluated how the number of the nearest neighbors affected the time cost. The results are shown in Figs. 3(a), 3(b), 4(a), and 4(b). We present the time costs of both kNN-BRT and Spilltree-BRT algorithms. We found that Spilltree-BRT performed better when the number of the nearest neighbors was small. This is also because it backtracked fewer times to the parent nodes when fewer neighbors were needed.



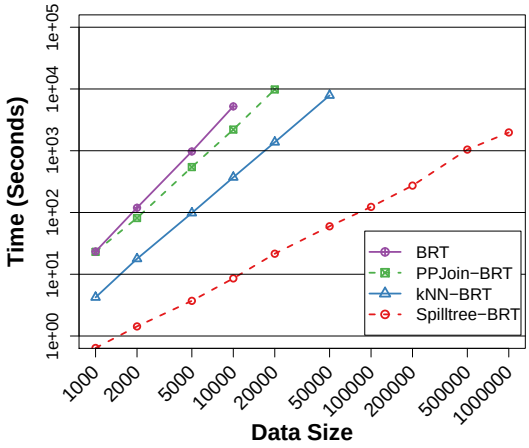
(a) Preprocessing time for query data.



(b) Building time for query data.

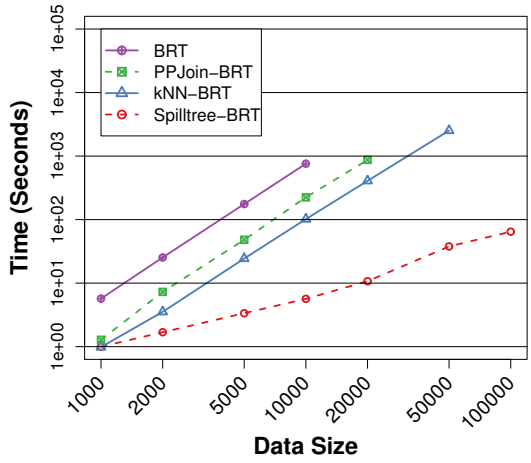


(c) Preprocessing time for news data.

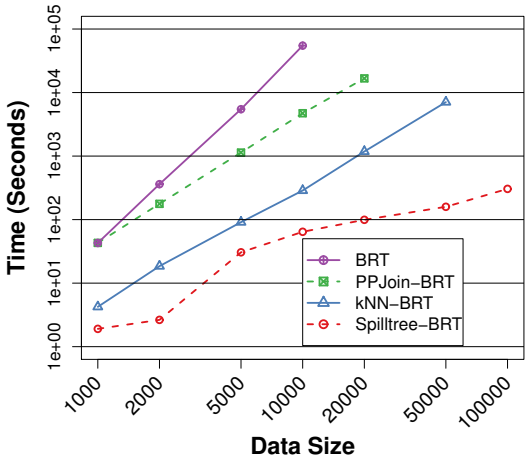


(d) Building time for news data.

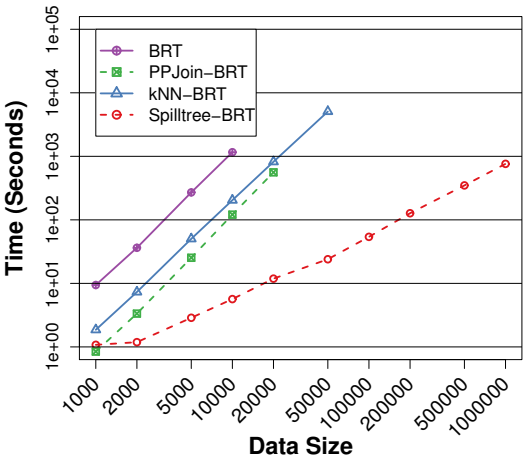
Figure 1: Time cost comparison of different algorithms based DCM distribution.



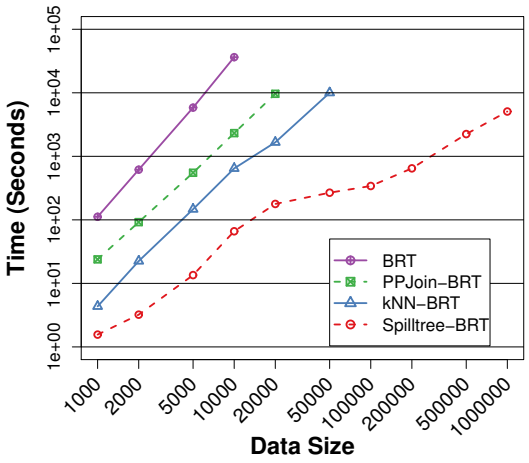
(a) Preprocessing time for query data.



(b) Building time for query data.

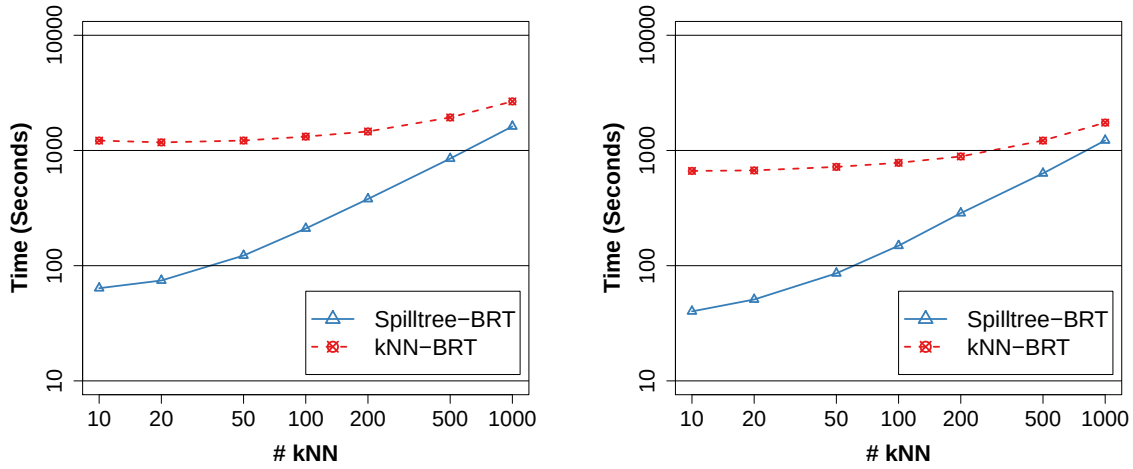


(c) Preprocessing time for news data.



(d) Building time for news data.

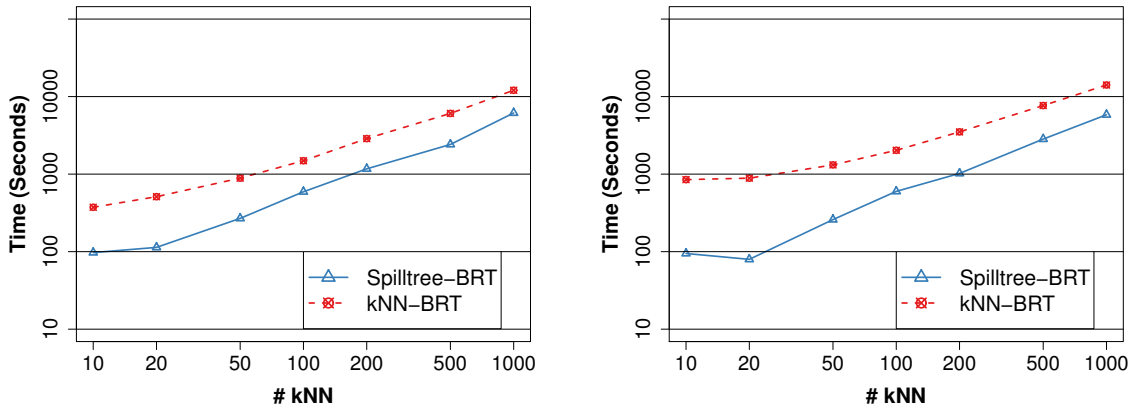
Figure 2: Time cost comparison of different algorithms based on vMF distribution.



(a) Time cost comparison of k nearest neighbors for query data.

(b) Time cost comparison of k nearest neighbors for news data.

Figure 3: The impact of different parameters and settings for kNN based methods (10,000 data samples).



(a) Time cost comparison of k nearest neighbors for query data.

(b) Time cost comparison of k nearest neighbors for news data.

Figure 4: The impact of different parameters and settings for kNN based methods (10,000 data samples).