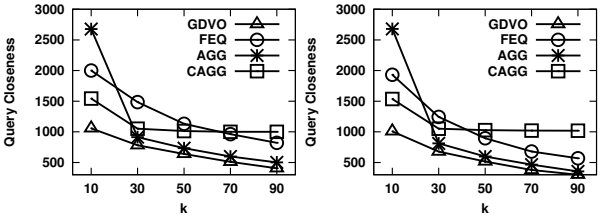
Algorithm 2 Computing $\triangle_g(x|S)$

Require: A tree T, a query I, a summary set S, a node $x \in \mathcal{V}$.
Ensure: $\triangle_g(x|S)$.
1: $S' \leftarrow S \cup \{x\}$;
2: Compute $\Phi_{S'}(x) = \{y \in \text{dec}(x) : \text{smy}_{S'}(y) = \text{rep}_x(y)\}$;
3: **if** anc (x) $\cap S \neq \emptyset$ **then**
4: Let $z \in S$ be the nearest ancestor of x;
5: $\triangle_g(x|S) = \sum_{y \in \Phi_{S'}(x)} (\text{rep}_x(y) - \text{rep}_z(y))$;
6: **else**
7: $\triangle_g(x|S) = \sum_{y \in \Phi_{S'}(x)} \text{rep}_x(y)$;
8: **return** $\triangle_g(x|S)$;



(a) physician (b) nurses

Figure 2: Quality evaluation on *physician* and *nurses* data

Complexity Analysis. The overall time complexity of Algorithm 1 is $O(n^2k)$ time in worst cases. The space complexity is $O(n)$.

6 EXPERIMENTS

In this section, we test our algorithms in experiments.

Datasets. We use a real-world dataset of tree T containing hierarchical terminologies that are extracted from the Medical Entity Dictionary (MED) [4]. The tree contains 4,226 nodes. In addition, we use two datasets of $\mathcal{I}$, where one dataset *physician* contains the information about how physicians query online knowledge resources, and the other dataset *nurses* contains the query information of nurses. These two datasets contain 2,425 records and 2,034 records, respectively. Each record consists of a MED term with a frequency count of its occurrence in the log file.

Methods Compared. To evaluate our algorithm GVDO, we evaluate and compare three algorithms – FEQ, AGG, and CAGG. Here, FEQ is a baseline approach, which selects k nodes with the highest frequencies [4]. The algorithm AGG picks a set of k nodes with the highest aggregate frequencies, where the aggregate frequency of a node x is defined as $AF(v) = \sum_{y \in \text{dec}(x)} \text{feq}(y)$. CAGG is a variant method of AGG using another metric of contribution ratio. For a node x, the contribution ratio of x is defined by $R(v) = \frac{AF(x)}{AF(y)}$ where y is the parent of x. Given a ratio threshold θ, CAGG selects the k nodes that have the highest aggregate frequencies and the contribution ratio no less than θ. We set $\theta = 0.4$ by following [4]. For all methods, we set the parameter $k = 30$ by default.

Evaluation Metrics. To evaluate the quality of summary result S found by all algorithms, we randomly generate a set of query nodes Q following the frequency distribution of input nodes, and measure the closeness distance between query Q and summary S, denoted by $D(Q, S) = \sum_{q \in Q} \min_{x \in S} \text{dist}_T(q, x)$, where $\text{dist}_T(q, x)$ is the number of edges connecting q and x in tree T. The smaller is $D(Q, S)$, the better is the summary.

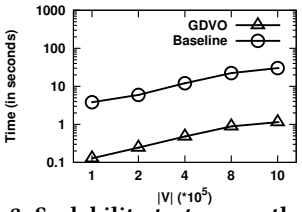


Figure 3: Scalability test on synthetic data

Quality Evaluation. Figures 2(a) and 2(b) show the quality evaluation on *physician* and *nurses* data by all algorithms. All approaches achieve smaller closeness distance with the increased k. Our approach GVDO is a clear winner of all competitors. It significantly outperforms the other methods for a smaller k, which is a great help to shrink large datasets for data summarization and visualization. The similar results can be observed in Figure 2(b).

Scalability Test. In this experiment, we evaluate the scalability of GVDO by varying the size of tree $|\mathcal{V}|$. We randomly generate 5 trees with size varying from 10^5 to 10^6. In addition, to verify the efficiency of computing $\triangle_g(x|S)$ by Algorithm 2, we compare one approach Baseline that follows Algorithm 1 by computing $\triangle_g(x|S)$ from scratch. The results of running time are shown in Figure 3. As we can see, GVDO is scalable very well with the increased size of tree nodes $|\mathcal{V}|$. Meanwhile, GVDO is much more efficient than Baseline, indicating the efficient strategy of Algorithm 2.

7 CONCLUSION

In this paper, we study the problem of ontology-based graph summary for visualization, and propose an efficient greedy algorithm with quality guarantee. Experiments on real-world datasets demonstrate the superiority of our proposed algorithm.

ACKNOWLEDGMENTS

This work was supported by the Hong Kong General Research Fund (GRF) Project Nos. HKBU 12200917, 12232716, 12200114, 12244916, and NSFC Grant No. 61672161.

REFERENCES

[1] I. Catallo, E. Ciceri, P. Fraternali, D. Martinenghi, and M. Tagliasacchi. Top-k diversity queries over bounded regions. *ACM Transactions on Database Systems (TODS)*, 38(2):10, 2013.
[2] I. F. Ilyas, G. Beskales, and M. A. Soliman. A survey of top-k query processing techniques in relational database systems. *ACM Computing Surveys (CSUR)*, 40(4):11, 2008.
[3] X. Jing, J. Cimino, et al. A complementary graphical method for reducing and analyzing large data sets. *Methods of information in medicine*, 53(3):173–185, 2014.
[4] X. Jing and J. J. Cimino. Graphical methods for reducing, visualizing and analyzing large data sets using hierarchical terminologies. In *AMIA Annual Symposium Proceedings*, volume 2011, page 635, 2011.
[5] R.-H. Li, J. X. Yu, X. Huang, H. Cheng, and Z. Shang. Measuring robustness of complex networks under mvc attack. In *CIKM*, pages 1512–1516, 2012.
[6] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions-i. *Mathematical Programming*, 14(1):265–294, 1978.
[7] S. Noel and S. Jajodia. Managing attack graph complexity through visual hierarchical aggregation. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*, pages 109–118, 2004.
[8] L. Qin, J. X. Yu, and L. Chang. Diversifying top-k results. *PVLDB*, 5(11):1124–1135, 2012.
[9] Y. Tian, R. A. Hankins, and J. M. Patel. Efficient aggregation for graph summarization. In *SIGMOD*, pages 567–580, 2008.
[10] Y. Wu, J. Gao, P. K. Agarwal, and J. Yang. Finding diverse, high-value representatives on a surface of answers. *PVLDB*, 10(7):793–804, 2017.
[11] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.