Panther: Fast Top-k Similarity Search on Large Networks

Jing Zhang1, Jie Tang2, Cong Ma3, Hanghang Tong2, Yu Jing1, Juanzi Li1

1Department of Computer Science and Technology, Tsinghua University
2School of Computing, Informatics, and Decision Systems Engineering, Arizona State University

Goal: Develop a Fast Top-k Similarity Algorithm for Large Networks

- Who are similar with Barabási?

- Similar authors in Aminer

- Efficiency Performance

Data set: Tencent Weibo: $|V| = 0.3$ billion, $|E| = 6$ billion.

Extract 11 different Tencent networks.

Our Approach: Panther

Pantherps

Path Similarity

Two vertices are similar if they frequently appear on the same paths.

\[ S_{2p}(v_1, v_2) = \frac{\sum_{p \in \text{T-paths in } G} w(p)}{\sum_{p \in \text{T-paths in } G} w(p)} \]

Uniformly sample a starting node.

Random walk according to the transition probability:

\[ t_j = \frac{1}{\sum_{i \in \mathbb{N}(v_j)} \delta(i)} \]

Random walk according to

Panther vs

Use top-D path similarities calculated by Pantherps to represent a vector:

\[ \theta(v_j) = (S_{2p}(v_1, v_{(1)}), S_{2p}(v_1, v_{(2)}), \ldots, S_{2p}(v_1, v_{(D)})) \]

Build kd-tree based on the Euclidean distance between any vectors.

\[ S_{vs}(v_j, v_k) = \frac{1}{|E|} \frac{|\theta(v_j) - \theta(v_k)|}{\max(|\theta(v_j)|, |\theta(v_k)|)} \]

Vector Similarity

The probability distributions of a vector linking to all other vertices are similar if their topology structures are similar.

\[ S_{vs}(u, v) = 0.27 > S_{vs}(u, v) = 0.16 \]

Time complexity

Pantherps: $O(RTc + NdT)$, Panthervs: $O(RTc + NdT + Nc)$

Experiments

Data set: Tencent Weibo: $|V| = 0.3$ billion, $|E| = 6$ billion.

- Extract 11 different Tencent networks.

Experimental Design

Efficiency Performance

Accuracy Performance

Parameter Analysis

Efficiency Performance

- Working with billions of edges.

- Stable when $T=5$ and $D=50$, refer to our paper for details.

- Stable when $T=5$ and $D=50$, refer to our paper for details.

- Co-author networks: $|V| = 3K$, $|E| = 7K$.

- Extract 11 different Tencent networks.

- Work with billions of edges.

- Stable when $T=5$ and $D=50$, refer to our paper for details.

- Co-author networks: $|V| = 3K$, $|E| = 7K$.

- Extract 11 different Tencent networks.

- Work with billions of edges.