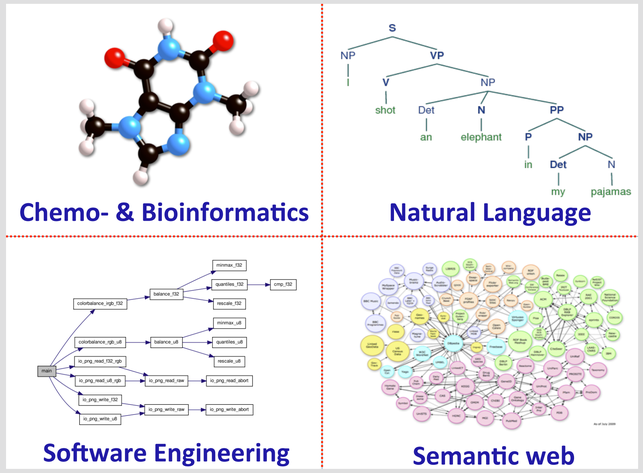
A Fast Kernel for Attributed Graphs

Yu Su[1], Fangqiu Han[1], Richard E. Harang[2], and Xifeng Yan[1]
[1]University of California, Santa Barbara, [2]U.S. Army Research Lab

Introduction

- A graph kernel defines a similarity measure over graphs, a core problem of graph mining. Graph kernels have been widely used in various application domains.

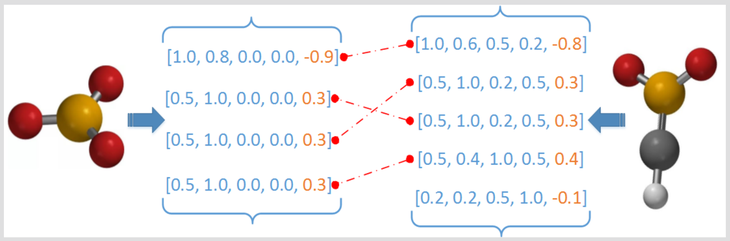


- Trends and core challenges in the big data era:
  (1) Increasing graph size calls for more efficient methods.
  (2) Richer graph attributes calls for more versatile methods.

- We propose a linear-time graph kernel which can handle both categorical and numerical attributes. Extensive experiments on both synthetic and real-world graph datasets show promising performance in both accuracy and efficiency.

Method: Descriptor Matching Kernel

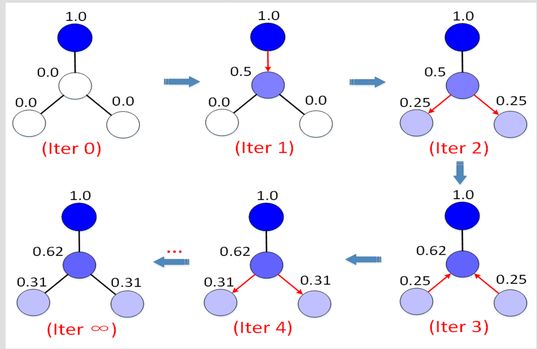
Overview



Method: Descriptor Matching Kernel (Cont'd)

Description Generation via Propagation

[Intuition] A descriptor needs to capture both the attributes and the neighborhood information of a node. Similar nodes should have similar descriptors.



(1) Initialization:

$$A_i^{(0)}(v) = \begin{cases} 1 & \text{if } \mathcal{L}_c(v) = l_i, \\ 0 & \text{otherwise;} \end{cases}$$

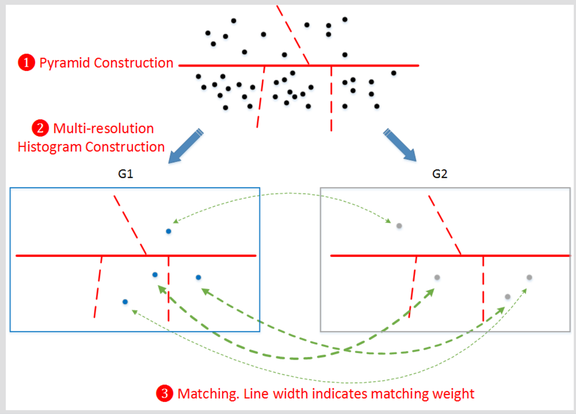
(2) Updating:

$$A_i^{(r+1)}(v) = \begin{cases} 1 & \text{if } A_i^{(r)}(v) = 1, \\ 1 - \prod_{u \in \mathcal{N}(v)} (1 - \eta A_i^{(r)}(u)) & \text{otherwise,} \end{cases}$$

$$for\ i = 1, \ldots, L,\ 0 \le r < h.$$

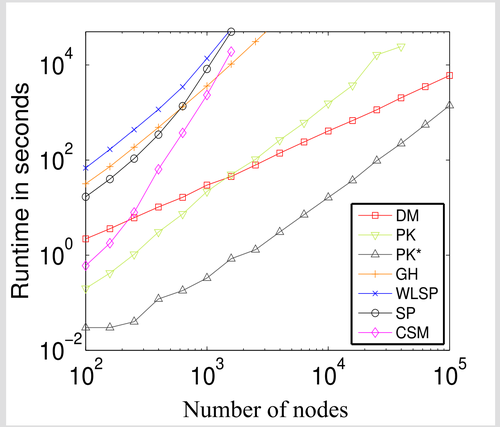
Descriptor Matching via Pyramid Matching Kernel

[Steps] (1) Hierarchical partitioning of the descriptor space based on data distribution. (2) Representing each graph as a multi-resolution histogram. (3) Bottom-up matching.



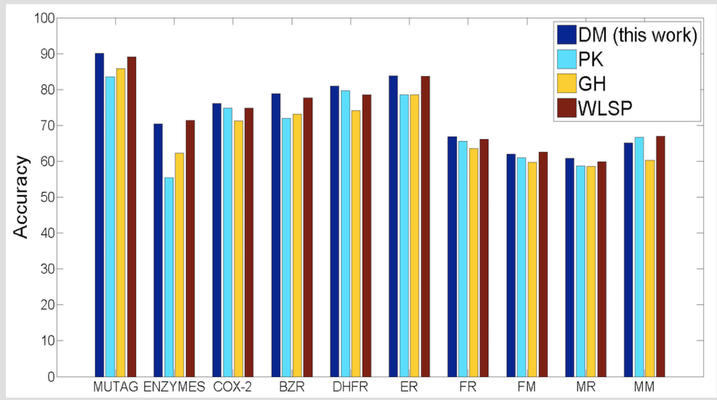
Experiments

Efficiency on Synthetic Graphs



Our DM kernel is only slower than the PK kernel, and is orders of magnitude faster than all the other kernels on large graphs with thousands of nodes.

Accuracy on Real-world Graphs



DM is among the best in 9 out of the 10 tested datasets. Under student's t test at p=0.05, DM is significantly better than PK on 8 datasets.

Conclusion

- We proposed a graph kernel that can handle both categorical and numerical attributes, while achieving a runtime linear *w.r.t.* graph size.
- Experiments on synthetic and real-world graphs showed competitive performance in both accuracy and efficiency.