CS 591.03
Introduction to Data Mining
Instructor: Abdullah Mueen

LECTURE 9: GRAPH MINING

Graph Similarity

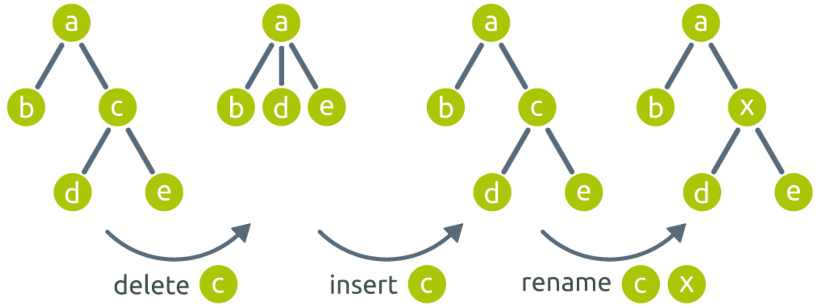
Edit distance/graph isomorphism:
- Tree Edit Distance

Feature extraction
- IN/out degree
- Diameter

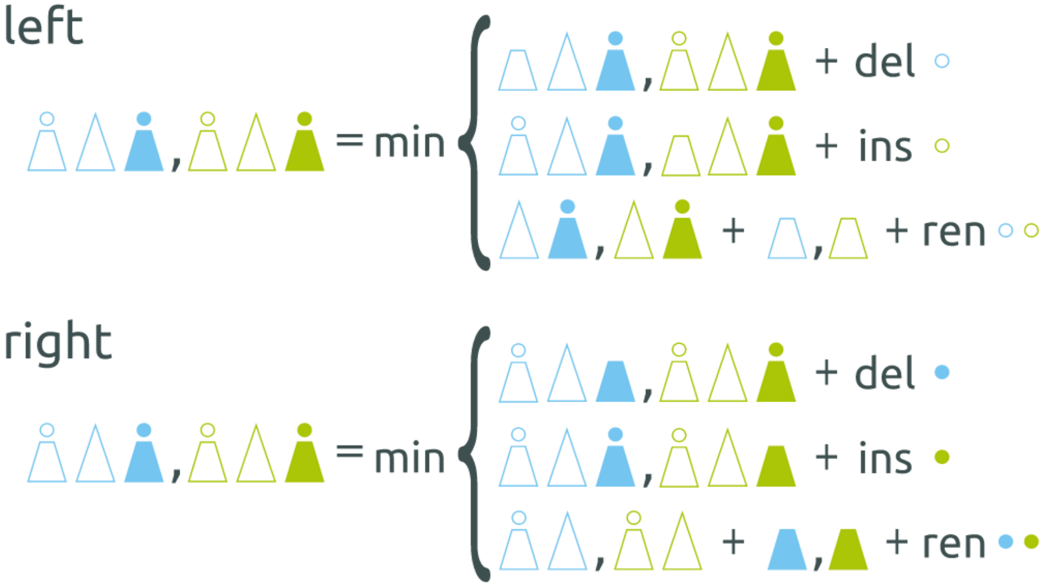
Iterative methods
- SimRank

Edit Distance

Three Operations



Tai's algorithm runs in $O(m^3n^3)$ time and space for trees with m and n nodes respectively.



Diameter

Largest Shortest path in the graph.

$$\text{shortestPath}(i,j,0) = w(i,j)$$

$$\text{shortestPath}(i,j,k{+}1) = \min(\text{shortestPath}(i,j,k), \text{shortestPath}(i,k{+}1,k){+}\text{shortestPath}(k{+}1,j,k))$$

```
let dist be a |V| × |V| array of minimum distances initialized
to ∞ (infinity)
for each vertex v
   dist[v][v] ← 0
for each edge (u,v)
   dist[u][v] ← w(u,v)  // the weight of the edge (u,v)
for k from 1 to |V|
   for i from 1 to |V|
      for j from 1 to |V|
         if dist[i][j] > dist[i][k] + dist[k][j]
              dist[i][j] ← dist[i][k] + dist[k][j]
          end if
```

http://en.wikipedia.org/wiki/Floyd%E2%80%93Warshall_algorithm

Simrank

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b))$$

For a node v in a graph, we denote by I(v) and O(v) the set of in-neighbors and out-neighbors of v, respectively.

1. A solution $s(*, *) \in [0, 1]$ to the n^2 SimRank equations always exists and is unique.
2. Symmetric
3. Reflexive

Ranking Nodes

Page Rank

$$PR(A) = (1-d) + d\ (PR(T1)/C(T1) + \ldots + PR(Tn)/C(Tn))$$

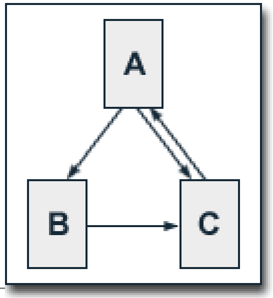
PR(A) is the PageRank of page A,

PR(Ti) is the PageRank of pages Ti which link to page A,

C(Ti) is the number of outbound links on page Ti and

d is a damping factor which can be set between 0 and 1.

Example



PR(A) = 0.5 + 0.5 PR(C)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

These equations can easily be solved. We get the following PageRank values for the single pages:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

HITS: Hyperlink-Induced Topic Search

Iterate(G,k)

G: a collection of n linked pages

k: a natural number

Let z denote the vector $(1, 1, 1, \ldots, 1) \in \mathbf{R}^n$.

Set $x_0 := z$.

Set $y_0 := z$.

For $i = 1, 2, \ldots, k$

Apply the $\mathcal{I}$ operation to (x_{i-1}, y_{i-1}), obtaining new x-weights x'_i.

Apply the $\mathcal{O}$ operation to (x'_i, y_{i-1}), obtaining new y-weights y'_i.

Normalize x'_i, obtaining x_i.

Normalize y'_i, obtaining y_i.

End

Return (x_k, y_k).