On measuring of similarity between tree nodes

Gleb B. Sologub

Department of Applied Mathematics and Physics

Moscow Aviation Institute (State Technical University), Volokolamskoe sh. 4

Moscow, 125993, Russia, http://www.mai.ru

glebsologub@ya.ru

Abstract

For different methods of information retrieval, data mining and related areas, basic task is to calculate similarity between data entries. Tree structures are used to represent various types of hierarchical data. Examples include different ontologies, catalogs, genealogies, XML documents, language corpuses, etc.

In our work on intelligent tutoring and testing systems we need to evaluate similarity between the questions of a test in order to predict answer scores. We use tree data structures for domain modeling. Nodes of a tree represent themes or subjects; leaves represent questions. So, the main goal of our study is to develop effective and accurate measure of similarity between tree leaves.

Here we present a survey of similarity measures between vertices of a graph. We describe distance-based and structural equivalence measures. It is demonstrated that most of them degenerate if applied directly to the tree nodes, and, especially, tree leaves. Adjusted path-based similarity measure is proposed as well as a new method for representing tree nodes as binary vectors that is based on using of an ancestor matrix. It is shown that application of ordinary similarity measures to this representation gives desired non-trivial results.

Notation

n	number of nodes
v_i	nodes (vertices)
t	root node of a tree
q_i	leaves of a tree
t_i	parent (non-leaf) nodes
lca_{ij}	lowest common ancestor of vertices v_i and v_j
$l(v_i,v_j)$	length of the shortest path between vertices v_i and v_j
n_{ij}	number of common neighbors of vertices v_i and v_j
$\mathbf{A}$	adjacency matrix
a_{ij}	elements of adjacency matrix
A_i	rows of adjacency matrix
$\mathbf{I}$	identity matrix
$\tilde{\mathbf{A}}$	ancestor matrix
$\mathbf{C}$	extended ancestor matrix, which is $\mathbf{I} + \tilde{\mathbf{A}}$
k_i	degree of ith vertex (number of neighbors)
$\mathbf{D}$	diagonal degree matrix with elements $d_{ii} = k_i$
$\mathbf{L}$	Laplacian matrix, which is $\mathbf{D} - \mathbf{A}$
$\mathbf{\Gamma}$	Moore-Penrose inverse of the Laplacian matrix $\mathbf{L}$
$x, y, \ldots$	vectors
$\|x\|$	norm of vector x
(x, y)	dot product of vectors x and y
θ	angle between vectors
$\mathrm{cov}(x, y)$	covariation of vectors x and y
σ_x	standard deviation of components of vector x
$\bar{x}$	sample mean of components of vector x
$M, N, \ldots$	sets

Preliminaries

A tree is a connected undirected simple graph with no cycles. We consider a rooted tree, which has a root node and leaves. Any leaf has only one neighbor, which is its parent. Any two nodes of a tree are connected by a unique simple path, which is the shortest path between them. Any two nodes of a tree have the lowest common ancestor and $l(v_i, v_j) = l\left(v_i, lca_{ij}\right) + l\left(v_j, lca_{ij}\right)$.

Note that $a_{ij} = a_{ji} = \{0 \text{ or } 1\}$ and $a_{ij}^2 = a_{ij}$ for all i, j. Also, it is obvious that $n_{ij} = (A_i, A_j) = \sum_k a_{ik} a_{kj}$ and $k_i = n_{ii} = (A_i, A_i) = \|A_i\|^2 = \sum_k a_{ik}$.

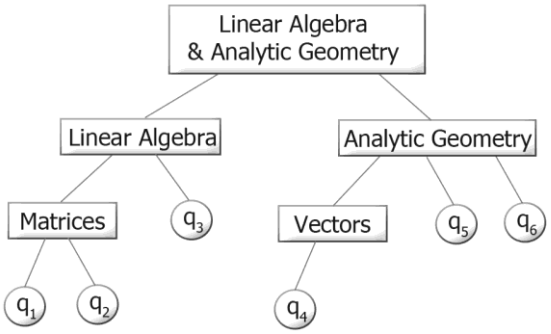
Then, in the case of tree leaves $k_i = 1$ and $n_{ij} = \begin{cases} 1, \text{ if } q_i \text{ and } q_j \text{ have the same parent;} \\ 0, \text{ otherwise.} \end{cases}$

If $d(x,y)$ is a distance between objects of any kind, e.g. vectors x and y, then their similarity could be measured as $s(x, y) = \frac{1}{1 + d\left(x, y\right)}$.

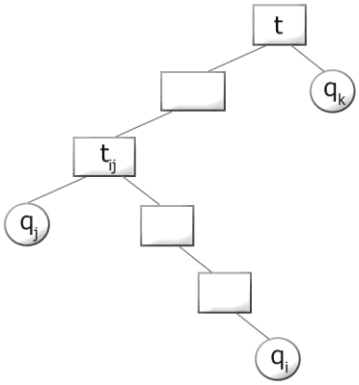
Distances on vertices

Distance	Shortest path, l	Resistance distance, Ω_{ij}	Adjusted shortest path, l_a
Definition	$l(v_i, v_j)$	$\Gamma_{ii} + \Gamma_{jj} - \Gamma_{ij} - \Gamma_{ji}$	$\frac{l(v_i, v_j)}{1 + l(lca_{ij}, t)}$
Similarity measure	$\frac{1}{1 + l(v_i, v_j)}$		$\frac{1 + l(lca_{ij}, t)}{1 + l(lca_{ij}, t) + l(v_i, v_j)}$
Drawback	No difference between similarities of node pairs located at different depths*	In the case of a tree is equal to the shortest path	l_a is not a metric**

* $l(q_1, q_2) = l_a(q_5, q_6)$:



** $l_a(q_i,q_k) > l_a(q_i,q_j) + l_a(q_j,q_k)$:



However, we can define a proper metric as $\tilde{l}_a(v_i, v_j) = \frac{l(v_i, v_j)}{1 + l(lca_{ij}, t) + l(v_i, v_j)}$.

Structural equivalence measures

Measure	Euclidean distance, ρ_E	Tanimoto coefficient, S_T	Cosine similarity, σ_{ij}	Pearson correlation coefficient, r_{ij}
Origin	$\|x-y\|$	$\frac{\|M\cap N\|}{\|M\|+\|N\|-\|M\cap N\|}$	$\cos\theta$	$\frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y}$
Definition on vectors	$\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$	$\frac{(x,y)}{\|x\|^2+\|y\|^2-(x,y)}$	$\frac{(x,y)}{\|x\|\cdot\|y\|}$	$\frac{\sum_k(x_k-\bar{x})(y_k-\bar{y})}{\sqrt{\sum_k(x_k-\bar{x})^2}\sqrt{\sum_k(y_k-\bar{y})^2}}$
On rows of adjacency matrix of a graph	$\sqrt{\sum_k a_{ik}^2+\sum_k a_{jk}^2-2\sum_k a_{ik}a_{kj}}$	$\frac{\sum_k a_{ik}a_{kj}}{\sum_k a_{ik}^2+\sum_k a_{jk}^2-\sum_k a_{ik}a_{kj}}$	$\frac{\sum_k a_{ik}a_{kj}}{\sqrt{\sum_k a_{ik}^2}\sqrt{\sum_k a_{jk}^2}}$	$\frac{\sum_k\left(a_{ik}-\frac{1}{n}\sum_m a_{im}\right)\left(a_{jk}-\frac{1}{n}\sum_m a_{jm}\right)}{\sqrt{\sum_k\left(a_{ik}-\frac{1}{n}\sum_m a_{im}\right)^2}\sqrt{\sum_k\left(a_{jk}-\frac{1}{n}\sum_m a_{jm}\right)^2}}$
Using degrees and counts of common neighbors	$\sqrt{k_i+k_j-2n_{ij}}$	$\frac{n_{ij}}{k_i+k_j-n_{ij}}$	$\frac{n_{ij}}{\sqrt{k_ik_j}}$	$\frac{n_{ij}n-k_ik_j}{\sqrt{k_in-k_i^2}\sqrt{k_jn-k_j^2}}$
For leaves of a tree	$\sqrt{2-2n_{ij}}$	$\frac{n_{ij}}{2-n_{ij}}$	n_{ij}	$\frac{n_{ij}n-1}{n-1}$
For leaves having the same parent	0	0.5	1	1
For leaves having different parents	$\sqrt{2}$	0	0	$-\frac{1}{n-1}$
On rows of extended ancestor matrix of a tree	$\sqrt{\|C_i\|^2+\|C_j\|^2-2(C_i,C_j)}$	$\frac{(C_i,C_j)}{\|C_i\|^2+\|C_j\|^2-(C_i,C_j)}$	$\frac{(C_i,C_j)}{\|C_i\|\|C_j\|}$	$\frac{\frac{n(C_i,C_j)}{\|C_i\|\|C_j\|}-\|C_i\|\|C_j\|}{\sqrt{n-\|C_i\|^2}\sqrt{n-\|C_j\|^2}}$
Path-based representation for tree nodes	$\sqrt{l(v_i,v_j)}$	$\frac{1+l(lca_{ij},t)}{1+l(lca_{ij},t)+l(v_i,v_j)}$	$\frac{1+l(lca_{ij},t)}{\sqrt{(1+l(v_i,t))(1+l(v_j,t))}}$	$\frac{n(1+l(lca_{ij},t))-(1+l(v_i,t))(1+l(v_j,t))}{\sqrt{(n(1+l(v_i,t))-(1+l(v_i,t))^2)(n(1+l(v_j,t))-(1+l(v_j,t))^2)}}$

Proposed representation of tree vertices

The ancestor matrix $\tilde{\mathbf{A}}$ of a graph is defined as a square matrix where an element $\tilde{a}_{i,j}$ is set to 1 if the *j*th vertex is an ancestor of the *i*th vertex, and 0 otherwise.

We propose to use rows of $\mathbf{C} = \mathbf{I} + \tilde{\mathbf{A}}$ matrix as binary vectors for measuring of distances and similarity between vertices of a tree.

This matrix is less sparse and gives us more sophisticated measure. Also, it yields the highest value of similarity only for identical vertices, as opposed to adjacency matrix. This behavior is preferred in most cases.

Consider the sets $P_i = \{v_i, t_{i_1}, t_{i_2}, ..., lca_{ij}, t_{k_1}, t_{k_2}, ..., t\}$ and $P_j = \{v_j, t_{j_1}, t_{j_2}, ..., lca_{ij}, t_{k_1}, t_{k_2}, ..., t\}$, where t is the root of tree T, lca_{ij} is the lowest common ancestor of vertices v_i and v_j of T, t_{k_p} are their other common ancestors; t_{i_l}, t_{j_m} are the other ancestors of given vertices v_i and v_j, respectively.

Another result of our approach is that we can establish following connections between rows of extended ancestor matrix $\mathbf{C}$ and paths related to respective vertices:

$$\|C_i\|^2 = |P_i| = 1 + l(v_i, t) = 1 + l(v_i, lca_{ij}) + l(lca_{ij}, t),$$

$$(C_i, C_j) = |P_i \cap P_j| = |\{lca_{ij}, t_{k_1}, t_{k_2}, ..., t\}| = 1 + l(lca_{ij}, t).$$

These equations allow us to derive path-based expressions for nearly any kind of measures defined on vectors.

Possible applications

- **Text retrieval**: detection of similar documents that is based on the concept of semantic similarity, which uses language corpus statistics and lexical taxonomy.
- **Version control systems**: tree comparison methods, e.g. consensus methods that are based on computing of similarity between tree nodes.
- **Content-based image retrieval**: methods based on distance between images represented by a tree structure resulting from a recursive image partition.
- **Student diagnosis**: methods that use modeling of knowledge contained in taxonomies.
- **Phylogenetic trees matching**: methods based on using of similarity measure induced by tree structure.
- **Social network service**: friend searching based on address, interests taxonomy, professional niche, etc.
- **Recommender systems**: content-based recommendation methods.