Update Network by Merging Nodes

Description

'Update Network by Merging Nodes' lets you combine two or more duplicate nodes together into a single node. To do this, you must provide the algorithm a network, and a Merge Table for that network. See section 5.1.4.2 of the Sci2 tutorial for more on Merge Tables. You will also need to provide a special . properties file as a parameter if the node and edge attributes of the network are important to you.

The .properties file for this algorithm lets you specify how the attributes of two nodes or two edges should be combined to form the attributes of the merged node or edge. Let's use the file 'mergeIsiAuthors.properties', found in the /sampledata/isi/properties/ folder of your Sci2 tool, as an example.

```
node.numberOfWorks = numberOfWorks.sum
node.timesCited = timesCited.sum
edge.numberOfCoAuthoredWorks = numberOfCoAuthoredWorks.sum
edge.weight = weight.ignore
```

The basic pattern for this file is :

[node OR edge].[attribute name] = [attribute name].[aggregation function]

The idea here is that for each set of nodes that are each others duplicate, we combine each of the old attributes values for those nodes using some mathematical function, in order to create the new attribute values of the final merged node. If an attribute is not mentioned, the algorithm will pick one of attribute values from the duplicate nodes to use in the resulting node. This is acceptable if the two values will always be the same, but not so much if they could differ.

[attribute name] is the name of the attribute we are going to operate on to create a new value for the final merged node. This can be the name of any of the node attributes in the network.

[aggregation function] determines how we aggregate, or combine, the values of the duplicate nodes into the value of the final merged node.

Your options for aggregation function are as follows (for numeric attributes only):

- sum - the sum of each duplicate node's attribute values. Example use: When you have two author nodes who are really the same author, and you want to combine the number of citations they have accumulated under both names.
- max - the maximum value of each duplicate node's attribute values. Example use: When you have two author nodes who are really the same author, which have two differing author ages, you might want to assume that the younger age was based on an old record, and keep the older age.
- min - the minimum value of each duplicate node's attribute values. Example use: Your guess is as good as mine. Seems like we should have it if we have max :)
- ignore - this is the same as note including the line in the .properties file at all.

Although this algorithm is for merging nodes, it also necessarily implies the merging of edges. For instance, if one node has an edge to two other nodes, and those two other nodes are merged to become one node, the edges from the first node to the two duplicate nodes will be combined into one edge.

Pros & Cons

This algorithm works, but could stand to have some more functions for dealing with text, e.g. a property to keep the longer of two lists of author names. This algorithm should eventually be replaced by functionality in the new database workflow.

See Also

The license could not be verified: License Certificate has expired! Generate a Free license now.

Usage Hints

The properties files for this algorithm are very similar to the properties files used in Extract Directed Network or Extract Co-Occurrence Network. For many pipelines you will need to create similar .properties files for both. Try creating one and adapting it for use in the other.