Lesson 11

Clustering and SimRank in Social Network Graph

.

Clustering in Social Network Graphs

- Methods of detecting communities by finding clustering and cluster coefficients
- A clustering coefficient is a metric for the likelihood that two associated vertices of a vertex are also associated with other vertices.

Higher Clustering Coefficient

- Indicates a greater association and cohesiveness

Connected Components

- Mean components of the datasets (represented by properties of vertices) connected together
- For example, finding student–teacher datasets, social network datasets, etc..

Spark GraphX algorithms

- For analyzing graphs. Connected components
- graph.connectedComponents().vertices method in SparkGraph
- Labels each connected component of the graph with an ID. Each connected component ID is ID of the lowest-numbered vertex.

Connected Component Objects

- Can approximate clusters
- GraphX contains an implementation of the algorithm in the ConnectedComponentsObject
- The clusters are found by discovering close-by connected components using closeness centrality metric

SimRank

- Similarity can be defined by properties of graph vertices
- For example course, subject, student, scientist, Java programmer, status, values, or any other salient characteristic
- Social network analysis of graphs computes *SimRank*

SimRank

- A metric for measuring similarity between vertices of the same type
- The computation starts from a vertex possessing specific property and path traversals through the edges search the similarities
- The vertices having similar properties are counted to the SimRank.

SimRank

- The counting continues till counts per unit traversals converge within a prefixed margin, say .001.
- SimRank converges to a value which is applicable for path traversals within, say geodesic distance, say up to 200. The computations are analogous to ones for PageRank as in Example 9.7.

Summary

We learnt:

- Clustering Coefficient
- Connected Components
- Spark GraphX Connected components functions
- SimRank

End of Lesson 11 on

Clustering and SimRank in Social Network Graph