

DM Assignment 11

9.3.2016

Adithya Abraham Philip

1PI13CS008

Overview

This document answers a question about the effect of epsilon and popularity threshold values on the SCAN algorithm.

Questions & Answers

1. What is the effect of varying the epsilon and/or the popularity threshold?


Ans: Assuming popularity is kept constant and greater than one:

Increasing the epsilon value should result in fewer core points and as a result more numerous, smaller clusters. In the extreme case (depending on the graph and popularity threshold) it can result in all points being outliers (epsilon becomes 1 and popularity threshold and graph structure prevent any core points from being formed).

Reducing epsilon (from 1) should initially result in more clusters (small in size) and lesser outliers, and further reducing it (as it moves closer to 0) should reduce the number of clusters but make each cluster larger. The number of outliers reduces throughout. Can result in a single large cluster in the extreme case.

Increasing the popularity threshold for a given epsilon makes it harder for nodes to become core points, consequently reducing the reach and formation of clusters, resulting in more outliers and more numerous, smaller clusters. Increasing it beyond a point will result in fewer, smaller clusters as less and less points qualify to be core points, eventually making all nodes outliers.

Reducing the popularity threshold causes a large number of core points to pop up, which depending on whether the epsilon value is high or low and the graph itself can result in a large number of small clusters (high epsilon) or a single large cluster (low epsilon).



However, the effect of varying epsilon and popularity threshold is best studied together, as summarised here:

Epsilon	Popularity Threshold	Effect
Low	Low	Tends toward one or two large clusters, fewer outliers.
Low	High	Tends toward more numerous smaller clusters, number of outliers stays more or less constant.
High	Low	Smaller, more numerous clusters, more outliers.
High	High	Almost every node is an outlier.

These are the general trends. However, the effects of both parameters also depend on the nature of the graph on which the algorithm is being run.