* Q11.4 – Compare the MaPle algorithm (Section 11.2.3) with the frequent closed itemset mining algorithm, CLOSET (Pei, Han, and Mao [PHM00]). What are the major similarities and differences? 519

The MaPle algorithm has some similarities and differences to the CLOSET algorithm. The major similarities between the two methods lie in the enumeration and search framework, as well as the pruning methods. Both methods employ a depth-first search and pruning based off of redundant supersets. The major differences lie in the fact that MaPle is acting on γ-pClusters to find the maximal values and guarantees completeness by enumerating all biclusters. However, this causes the algorithm to be extremely computationally intensive if the data set is large.

* Q11.6 – In a large spare graph where on average each node has a low degree, is the similarity matrix using SimRank still sparse? If so, in what sense? If not, why? Deliberate on your answer.

Yes, in a scenario with a large spare graph the similarity matrix via SimRank can still be sparse. This can occur, because in this situation each vertex connects to a small number of other vertices. Therefore, according to the equations called out in section 11.3.2, as well as the sparsity equation 11.38, this can lead to a spare similarity matrix.

* Q11.7 – Compare the SCAN algorithm (Section 11.3.3) with DBSCAN (Section 10.4.1). What are their similarities and differences?

The SCAN algorithm is specific to graphs and looks for clusters by finding well-connected data points. It utilizes structural-context similarity to find the similarity of two vertices over a normalized neighborhood. There is a similarity threshold, ε, that is used to determine if a point should be placed into a cluster. SCAN uses a core vertex to grow a cluster from, which is similar, but not identical, to the density-based cluster method in DBSCAN method, which uses core objects. DBSCAN also has a threshold value, but it is focused on a density of points value. Additionally, in SCAN, vertices can be defined within a cluster, a hub, or found to be an outlier. SCAN is fairly robust in terms of scalability, because the number of edges is the same scale as the number of vertices. SCAN looks for cuts in a graph to form vertices based on structural-context similarity, while DBSCAN looks for and finds clusters.

* Q11.2 (Instead of Q11.8 for Non-CS Students) – *AllElectronics* carries 1000 products, P1,…,P1000. Considers customers Ada, Bob, and Cathy such that Ada and Bob purchase three products in common P1, P2, and P3. For the other 997 products, Ada and Bob independently purchase seven of them randomly. Cathy purchases 10 products, randomly selected from the 1000 products. In Euclidean distance, what is the probability that *dist*(Ada,Bob) > *dist*(Ada,Cathy)? What if Jaccard similarity (Chapter 2) is used? What can you learn from this example?

By Euclidean Distance:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Dist(Ada,Bob) | 3.7 | 3.5 | 3.2 | 2.8 | 2.4 | 2 | 1.4 | 0 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Dist(Ada,Cathy) | 4.2 | 4 | 3.7 | 3.5 | 3.2 | 2.8 | 2.4 | 2 | 1.4 | 0 |

Based on Euclidean distance, the probability that *dist*(Ada,Bob) > *dist*(Ada,Cathy) is 4.7 x 10-9.

By Jaccard Similarity:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| J(Ada,Bob) | 0.18 | 0.25 | 0.33 | 0.43 | 0.50 | 0.67 | 0.82 | 1.00 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| J(Ada,Cathy) | 0.052 | 0.11 | 0.18 | 0.25 | 0.33 | 0.43 | 0.50 | 0.67 | 0.82 | 1.00 |

Based on Jaccard similarity, the probability that *J*(Ada,Bob) > *J*(Ada,Cathy) is 8.9 x 10-3.

As we can see from this example, there is a very small chance that two customers will choose common products when the pool of products to choose from is 1,000 products. In regards to the two methods that were used, we can see that the Jaccard similarity is normalized between values of 0 and 1; however, the Euclidean distance can be can value above 0. This means that values can differ drastically as the data set becomes larger and potentially more sparse.

