Abstract

In this final project used Cover Coefficient-Based Incremental Methodology (C2ICM) to do clustering in SMART dataset, ADI data and CISI data. The processes are preprocessing, making D-Matrix, calculating C-Matrix, calculating number of cluster, choosing seed documents, and clustering nonseed documents to seed documents chosen. If there is new document arrive in database, so the next process called incremental maintenance. All documents in database will be recomputed D-Matrix, C-Matrix, number of cluster, and seed power. If the old seed document is chosen again as seed, so the old cluster builded by this seed document will be released, but if the old documents is not chosen as seed anymore, so that the old cluster will be deleted, and the new clusters will be created depend on new seed documents. After getting cluster for all query, the next step is calculate cluster quality by using Silhouette Coefficient (SC) and the clustering time in seconds. From the experiments, many kinds of quality cluster resulted. The clustering time that resulted in this testing is linier with number of hitlist documents.

***Keywords**: incremental, cluster, seed, cover coefficient, silhouette coefficient, cluster quality*