Recuperação inteligente da informação
Relatório atividade 1.6

André Pacheco

Doutorado em Ciência da Computação
Programa de pós-graduação em informática
Universidade Federal do Espírito Santo

Conteúdo

1 Introdução

A atividade E-1.6 tem como objetivo a realização da clusterização da famosa base de dados da IRIS [1] utilizando o algoritmo k-means implementando dentro do framework aLine.

2 Resultados de clusterização

A clusterização da base de dados da Iris foi realizada e foi as metricas apresentadas por (Salton et al., 1975) foram replicadas na Tabela 1. Foram utilizdas 3 configurações clusters, com 2, 3 e 4 centros. A densidade média da base é de 0.956.
Como pode ser observado, o valor de IDF não influencia tanto os calculos por dois motivos, primeiro o aLine não utiliza o mesmo para a clusterização, segundo a base da iris é bem comportada, fazendo com os resultados seja bem próximos, como reportado na tabela.

[1] https://archive.ics.uci.edu/ml/datasets/iris

Tabela 1. Métricas para clusterização

	Cluster A K = 2		Cluster B K = 3		Cluster C K = 4	
Tipo de indexação	f_i^k	$f_i^k \times IDF$	f_i^k	$f_i^k \times IDF$	f_i^k	$f_i^k \times IDF$
Media de similaridade entre documentos e suas correspondentes centroides (x)	0.996	0.993	0.998	0.994	0.988	0.994
Media de similaridade entre as centroides e a centroide principal	0.977	0.977	0.997	0.977	0.981	0.981
Media de similaridade entre os pares de centroides (y)	0.456	0.456	0.623	0.623	0.713	0.713
Taxa y/x	0.458	0.459	0.624	0.626	0.714	0.717

Bibliografia

Salton, G., Wong, A., Yang, C.S., 1975. A vector space model for automatic indexing. Communications of the ACM 18, 613–620.