Recuperação inteligente da informação
Relatório atividade P-2

André Pacheco

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

Conteúdo

1 Introdução

A atividade P-2 tem como objetivo a realização da clusterização da base de dados aTribuna, UsielCarneiro e Iris utilizando o algoritmo k-means implementando dentro do framework aLine.

2 Resultados de clusterização

A clusterização da base de dados aTribuna e IRIS foram realizadas seguindo os mesmos procedimentos das atividades T-11 e E-1.5. As metricas apresentadas por (Salton et al., 1975) foram replicadas nas Tabelas 1, 2 e 3. Vale ressaltar que não houve tempo hábil para aplicar as bases para o software CLUTO devido a problemas de disponibilidade do servidor. Para cada base foram utilizadas 3 configurações de clusters, como apresentado nas tabelas.

Tabela 1. Métricas da clusterização - aTribuna

	Cluster A K = 17	Cluster B K = 21	Cluster C K = 27
Tipo de indexação	f_i^k	f_i^k	f_i^k
Media de similaridade entre documentos e suas correspondentes centroides (x)	0.132	0.410	0.355
Media de similaridade entre as centroides e a centroide principal	0.789	0.7653	0.865
Media de similaridade entre os pares de centroides (y)	0.660	0.656	0.776
Taxa y/x	4.999	1.597	2.183

Tabela 2. Métricas da clusterização - IRIS

	Cluster A K = 2	Cluster B K = 3	Cluster C K = 4
Tipo de indexação	f_i^k	f_i^k	f_i^k
Media de similaridade entre documentos e suas correspondentes centroides (x)	0.996	0.998	0.998
Media de similaridade entre as centroides e a centroide principal	0.977	0.997	0.981
Media de similaridade entre os pares de centroides (y)	0.456	0.623	0.713
Taxa y/x	0.458	0.624	0.715

As análises das bases aTribuna e IRIS já foram realizadas nas atividades anteriores já mencionadas. Na base de UsielCarneiro a configuração as três configurações obtiveram resultados semelhantes. Em todos casos, a media de similaridade entre os elementos de um mesmo cluster são próximas, sendo K = 5 e 8 levemente maior do que K = 3.

Tabela 3. Métricas da clusterização - UsielCarneiro

	Cluster A K = 3	Cluster B K = 5	Cluster C K = 8
Tipo de indexação	f_i^k	f_i^k	f_i^k
Media de similaridade entre documentos e suas correspondentes centroides (x)	0.7586	0.8273	0.8236
Media de similaridade entre as centroides e a centroide principal	0.9918	0.9950	0.9748
Media de similaridade entre os pares de centroides (y)	0.6556	0.7919	0.8355
Taxa y/x	0.8642	0.9572	1.014

Bibliografia

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