Recuperação inteligente da informação
Relatório atividade T-11

André Pacheco

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

Conteúdo

1 Introdução

A atividade T-11 tem como objetivo a realização da clusterização da base de dados aTribuna [1] utilizando o algoritmo k-means implementando dentro do framework aLine.

2 Resultados de clusterização

A clusterização da base de dados aTribuna foi realizada e as metricas apresentadas por (Salton et al., 1975) foram replicadas na Tabela **??**. Vale ressaltar, que não foi possível obter a clusterização para os 45 mil documentos pois a máquina utilizada não foi capaz de clusterizar em tempo hábil (aguardou-se 24h para 45k, 21 centros com 100 iterações do k-means e a máquina ainda estava processando). Portanto, foram utilizadas 5 mil documentos com 3 configurações clusters, com 17, 21 e 27 clusters e com 100 iterações do k-means.
Como pode ser observado, das três configurações de cluster, a com melhor resultado é a B. 21 clusters é o total de classes conhecidas da base, isso ajuda a explicar o resultados. Quando o número de cluster foi subtraido para 17, a taxa final chega a quase 5, a maior de todas. Por fim, é possível certificar a dificuldade em clusterizar essa base, pela diversidade e quantidade de documentos nela presente.

[1] http://www.inf.ufes.br/ elias/dataSets/aTribuna-21dir.tar.gz

Tabela 1. Métricas da clusterização

	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

Bibliografia

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