Universidade Federal de Pernambuco
Centro de Informática
Bacharelado em Ciência da Computação

AVALIAÇÃO TÉCNICAS DE AGRUPAMENTO K-MEANS BASEADAS EM MAPREDUCE

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Bruno Magalhães De Carvalho (bmc@cin.ufpe.br)
Orientador: Ricardo Bastos C. Prudencio (rbcp@cin.ufpe.br)
Área: Inteligência Computacional

1. CONTEXTO

Agrupamento (Clustering) é o processo de examinar uma coleção de "pontos", e agrupar os pontos em "clusters " de acordo com alguma medida de distância. O objectivo é que os pontos no mesmo aglomerado tenha uma pequena distância um do outro, enquanto que os pontos nos conjuntos diferentes estão a uma grande distância uns dos outros. [1] Agrupamento é uma classificação não-supervisionada de padrões (observações, itens de dados, ou vetores de características) em grupos. O problema de agrupamento foi abordada em muitos contextos e por pesquisadores em muitas disciplinas. Isso reflete seu amplo apelo e utilidade como um dos passos na análise exploratória de dados.[2]

Há várias aplicações que necessitam de agrupar um grande volume de dados, porém na maioria das abordagens propostos na literatura são técnicas que estão restrito a volumes razoavelmente pequenos de dados. [2] O algoritmo convergente K-means [2]e suas variações é amplamente usado, pois é fácil de implementar, roda muito rápido na prática e em tempo de execução no seu pior caso é exponencial.[3]

Muitos dos modernos algoritmos que processam e produzem grandes volumes de dados usam a técnica MapReduce [4]. MapReduce é um modelo de programação para processamento e geração de grandes volumes de dados. Programas que utilizam esse modelo, são automaticamente paralelizados e executados em um conjunto de máquinas de forma paralela e distribuídas altamente escalável. Todo o gerenciamento de partição de dados e tarefas, comunicação e tolerância a erros são abstraídos, tornando possível programadores sem experiência em Sistemas Distribuídos usá-los.

[5] proponhe um algoritmo para agrupamento baseado em K-means e sua implementação utiliza o modelo MapReduce (Parallel K-Means Clustering Based on MapReduce), os resultados experimentais mostram que PKMeans fornece bons resultados quanto velocidade, escala e tamanho no modelo MapReduce.

[6] proponhe outro algoritmo MRMk-means (Multiple Parallel MapReduce k-means Clustering with Validation and Selection) sua implementação usa o modelo MapReduce. A ideia principal é executar várias vezes para todos os valores de k (número de grupos) definido pelo usuário no intervalo [kmin, kmax], depois disso avaliar qual melhor partição resultante.

2. OBJETIVOS

O objetivo desse trabalho é implementar os algoritmos PKMeans[5] e MRMk-means[6], sobre a framework Hadoop, aplicá-los sobre uma grande base de dados e analisar os resultados. A avaliação será feita pelas medidas encontradas na literatura, para a comparação do desempenho de cada algoritmo.

3. CRONOGRAMA

Atividade	Mês															
	Abril				Maio				Junho				Julho			
Levantamento e estudo do material bibliográfico	■	■	■	■	■	■	■	■								
Implementação dos algoritmos				■	■	■	■	■	■							
Aplicação dos algoritmos aos dados e avaliação dos resultados								■	■	■	■	■				
Elaboração do Relatório											■	■	■	■		
Elaboração da apresentação												■	■	■		

4. POSSÍVEIS AVALIADORES

- Leandro Almeida (lma3@cin.ufpe.br)
- Renata Souza (rmcrs@cin.ufpe.br)

5. REFERÊNCIAS BIBLIOGRÁFICAS

[1] LESKOVEC, Jure; RAJARAMAN, Anand; D. ULLMAN, Jeffrey. Mining of Massive Datasets. 2014. DOI=http://www.mmds.org

[2] A. K. Jain, M. N. Murty, and P. J. Flynn. 1999. Data clustering: a review. ACM Comput. Surv. 31, 3 (September 1999), 264-323. DOI=http://dx.doi.org/10.1145/331499.331504

[3] David Arthur, Bodo Manthey, Heiko Röglin: k-Means has Polynomial Smoothed Complexity. *CoRR* abs/0904.1113 (2009)

[4] J. Dean and S. Ghemawat. MapReduce: Simplified data processing on large clusters. OSDI, pages 139–149, 2004.

[5] Weizhong Zhao,Huifang Ma, and Qing He. Parallel k-means clustering based on MapReduce. In *CloudCom*, pages 674–679, 2009.

[6] K. D. Garcia and M. C. Naldi, "Multiple Parallel MapReduce k-Means Clustering with Validation and Selection," Intelligent Systems (BRACIS), 2014 *Brazilian Conference* on, Sao Paulo, 2014, pp. 432-437. doi: 10.1109/BRACIS.2014.83