Algoritmo *K-Means* Paralelo com base no MapReduce para Mineração de dados agrícolas

Lays Helena Lopes Veloso[1], Luciano José Senger[1]

[1]Departamento de Informática – Universidade Estadual de Ponta Grossa (UEPG)
Caixa Postal 84030-900 – Ponta Grossa – PR – Brazil

lays.veloso@gmail.com, ljsenger@uepg.br

Abstract. *Clustering techniques are employed in applications in various fields of knowledge. The K-Means clustering algorithm is the most commonly used. However, the time spent in performing K-means can be considerable when large amounts of data are used. The aim of this work is to implement a MapReduce based parallel K-Means algorithm to run on a Hadoop cluster and improve the response time of data mining in agriculture. This algorithm will address deficiencies identified in other parallel implementations of K-Means. Its performance will be evaluated with respect to SpeedUp and ScaleUp by using large flux datasets from agricultural regions.*

Resumo. *Técnicas de agrupamento são empregadas em aplicações nas diversas áreas do conhecimento. O K-Means é o algoritmo de agrupamento mais comumente usado. No entanto, o tempo gasto para a execução do K-Means pode ser considerável quando grandes quantidades de dados são usadas. O objetivo deste trabalho é implementar o algorimo K-Means paralelo baseado no modelo MapReduce para ser executado em um cluster Hadoop e melhorar o tempo de resposta da mineração de dados agrícolas. Este irá tratar falhas identificadas em outras implementações paralelas do K-Means. Seu desempenho será avaliado com relação ao SpeedUp e ao ScaleUp a partir de experimentos usando grandes conjuntos de dados de fluxo de regiões agrícolas.*

1. Introdução

Com a modernização dos equipamentos de aquisição e transmissão de dados, as organizações têm investido em coletar uma massa de dados diária de observações, tais como medições de torres de fluxo e redes de sensores. Essa massa de dados vem sendo tratada pelo termo *Big Data* e traz como desafios, armazenar e processar os dados com tempo de resposta aceitável e com baixo custo.

A Mineração de Dados (MD) é um conjunto de técnicas que através do uso de algoritmos de Aprendizado de Máquina (AM), permitem extrair conhecimento a partir da identificação de padrões desconhecidos em dados e auxiliar à tomada de decisão [Witten e Frank 2005]. Conforme [Kudyba 2014] o uso de *Big Data* na MD pode melhorar à tomada de decisão, a partir do uso de todos os dados disponíveis, em vez de se limitar a pequenas parcelas dos dados. O MapReduce é o *framework* distribuído mais popular para a análise de *Big Data* [Sakr e Gaber 2014]. O Hadoop é um projeto *Open Source* para processamento distribuído que implementa o modelo MapReduce.

O objetivo deste trabalho é implementar o algoritmo *K-Means* paralelo com base no modelo MapReduce para ser executado em um *cluster* Hadoop e melhorar o tempo

de resposta da mineração de dados agrícolas. O algoritmo irá tratar falhas identificadas em outras implementações paralelas do *K-Means* que serão discutidas a seguir. O *K-Means* paralelo será avaliado com relação ao *SpeedUp*[6] e ao *ScaleUp*[7] a partir de experimentos usando grandes conjuntos de dados de fluxo de regiões agrícolas. Com essa tarefa se busca obter informações para o planejamento agrícola para controlar as emissões de gás carbônico (CO_2) na atmosfera, determinando o potencial de sequestro de carbono (C) do solo, e sua relação com as variáveis climáticas.

Os resultados encontrados serão avaliados com base na literatura e juntamente com um profissional da área da Agricultura Orgânica. A qualidade do agrupamento será avaliada utilizando os métodos *intra-cluster* e *inter-cluster*, como em [de Mello e Senger 2005].

2. *K-Means*

O *K-Means* consiste em reunir *n* amostras de dados em *k* grupos de maneira que as amostras em um mesmo grupo sejam similares entre si e diferentes daquelas em outros grupos [Sakr e Gaber 2014]. O algoritmo *K-Means* sequencial pode ser descrito em 4 passos:

1. Seleção de *k* amostras como centróides iniciais;
2. Atribuição de cada amostra ao centróide mais próximo com base em um critério de distância;
3. Cálculo de novos centróides através da média das amostras pertencentes ao mesmo centróide;
4. Os passos 2 e 3 são repetidos até convergir para uma solução ótima.

O *K-Means* pode tirar vantagem do paralelismo. As amostras podem ser distribuídas em cada processador e então atribuídas ao centróide mais próximo em paralelo [Dean 2014]. Algumas implementações paralelas do *K-Means* foram propostas, baseadas no MapReduce [ZHOU et al. 2011, Golghate e Shende 2014]. Em tais implementações, foram identificadas 2 falhas:

Ausência de tratamento de falta de dados: É comum nas bases de dados sequenciais casos em que as medidas não são realizadas por algum problema com os mecanismos de coleta ou gravação. Nosso algoritmo descarta as instâncias que apresentarem mais de 30% dos seus valores perdidos para evitar que esses registros prejudiquem o resultado final.

Falta de um método eficiente para a seleção dos centróides iniciais - A solução do *K-Means* é sensível aos centróides iniciais, que são geralmente selecionados de maneira aleatória. Nós iremos paralelizar uma técnica mais eficiente para a inicialização dos centróides de forma a melhorar a qualidade dos grupos finais.

3. Torres de Fluxo

Torres de fluxo utilizam a técnica de covariância de vórtices turbulentos ou Eddy Covariance (EC) para medir em longo prazo os fluxos de CO_2, água e outros nutrientes

[6] *SpeedUp* - Medida de ganho de desempenho de um algoritmo paralelo com relação a um algoritmo sequencial equivalente.
[7] *ScaleUp* - Medida da escalabilidade de um algoritmo, ou seja, a capacidade de um algoritmo lidar com porções crescentes de trabalho quando mais recursos estão disponíveis, de forma uniforme.

entre a atmosfera e os ecossistemas, florestais e agrícolas. Essas quantificações começaram a ser feitas em 1996 a fim de entender os controles sobre os fluxos de C [Cihlar et al. 2002]. O Fluxo pode ser definido como a quantidade de uma grandeza que passa através de uma superfície por unidade de tempo. O princípio geral das medições de EC é a covariância entre a concentração da grandeza de interesse e a velocidade vertical do vento [Burba 2013].

Para [Lichtfouse et al. 2011] a perda de carbono do solo merece uma atenção particular na Agricultura, pois controla diferentes fatores a longo prazo, tais como o CO_2 atmosférico, erosão e abastecimento de água e nutrientes. Com as descobertas de práticas agrícolas que podem diminuir os níveis de CO_2 na atmosfera, têm surgido diferentes estudos para observar a dinâmica do C a fim de descobrir o potencial de sequestro de C do solo em áreas com diferentes tratamentos [FAO 2011].

4. Implementação

4.1 Base de Dados

A base de dados a ser utilizada nos experimentos contém medições contínuas de fluxo do período de 2004 à 2011 coletadas em áreas agrícolas submetidas à diferentes tratamentos. Os dados pertencem a rede AmeriFlux[8] e foram obtidos em (http://ameriflux.ornl.gov/).

4.2 *K-Means* Paralelo

No modelo MapReduce o processamento é dividido em duas fases: *map* e *reduce*. Para isso é necessário especificar os passos da computação em duas funções respectivas. Desta maneira, o sistema de execução automaticamente paraleliza a aplicação através do *cluster* de computadores e cada iteração do *K-Means* é executada como um *job* MapReduce.

A aplicação inicia submetendo o *job*. Desta forma a base de dados é segmentada e suas partes são distribuídas em *Mappers* que são executados em paralelo. Os *Mappers* executam instâncias da função *map* em cada nó no *cluster* computacional. No *K-Means* os *Mappers* fazem a leitura dos centróides atuais, calculam a distância euclidiana entre os centróides e as amostras e atribuem cada amostra ao grupo com o centro mais próximo.

Como em [Zhao et al. 2009], uma função *combine* foi implementada para melhorar o desempenho do algoritmo. A função *combine* é executada no mesmo nó que a função *map* respectiva e tem como objetivo agrupar localmente as saídas da fase *map*. No *K-Means* implementado, a função *combine* faz a soma parcial das amostras atribuídas ao mesmo grupo na respectiva função *map* antes que a função *reduce* faça a soma total.

Por fim, a função *reduce* calcula e emite os valores dos novos centróides e o número de amostras atribuídas ao mesmo grupo para serem usados na próxima iteração do *K-Means*.

[8] AmeriFlux - Rede de torres de fluxo das Américas

4.3 *Cluster* Hadoop

O *cluster* Hadoop configurado para os experimentos possui uma máquina mestre com um processador Intel core i7-3537 de 2,50 GHZ com 2 núcleos, 4 processadores lógicos e 8,0 GB de memória RAM e seis máquinas escravas com processador Intel core i7-2600 de 3,4 GHz com 4 núcleos, 4 processadores lógicos e 4,0 GB de memória RAM.

5. Conclusões

Com este trabalho foi implementado o algoritmo de agrupamento *K-Means* paralelo com base no modelo MapReduce a fim de melhorar o tempo de resposta da mineração de dados agrícolas quando grandes conjuntos de dados são usados. A aplicação será executada em um *cluster* Hadoop com recursos computacionais de baixo custo. Será paralelizado um método eficiente para seleção dos centróides iniciais, que atualmente é feita de maneira aleatória. Além disso, a qualidade do agrupamento será avaliada com diferentes quantidades de grupos, para encontrar o número de grupos que minimiza as distâncias *intra-cluster* e maximiza as distâncias *inter-cluster*. O desempenho do algoritmo será avaliado e deverá apresentar *SpeedUp* e *ScaleUp* lineares. Desta maneira será possível fornecer resultados que efetivamente auxiliem a tomada de decisão agrícola para baixa emissão de CO_2 com alta velocidade de resposta.

Referências

Burba, G. (2013). Eddy Covariance Method for Scientific, Industrial, Agricultural and Regulatory Applications: A Field Book on Measuring Ecosystem Gas Exchange and Areal Emission Rates. LI-COR Biosciences.

Cihlar, J., Denning, A. S., and Gosz, J. R., editors (2002). Terrestrial Carbon Observation: The Ottawa Assessment of Requirements, Status and Next Steps. Number 2 in Environment and natural resources series. Food & Agriculture Org.

de Mello, R. F. and Senger, L. J. (2005). Automatic text classification using an artificial neural network. IFIP Advances in Information and Communication Technology, 172:215–238.

Dean, J. (2014). Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners. John Wiley & Sons.

FAO (2011). Organic agriculture and climate change mitigation. Technical report, FAO.

Golghate, A. A. and Shende, S. W. (2014). Parallel k-means clustering based on hadoop and hama. International Journal of Computing and Technology, 1.

Kudyba, S. (2014). Big Data, Mining, and Analytics: Components of Strategic Decision Making. CRC Press.

Lichtfouse, E., Hamelin, M., Navarrete, M., and Debaeke, P. (2011). Sustainable Agriculture, volume 2. Springer.

Sakr, S. and Gaber, M. (2014). Large Scale and Big Data: Processing and Management. CRC Press.

Witten, I. H. and Frank, E. (2005). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann series in data management systems, 2nd edition.

Zhao, W., Ma, H., and He, Q. (2009). Parallel k-means clustering based on mapreduce. In CloudCom 2009, volume 5931, pages 674–679. LNCS.

ZHOU, P., LEI, J., and YE, W. (2011). Large-scale data sets clustering based on mapreduce and hadoop. Journal of Computational Information Systems, 7:5956–5963.