Sistemas distribuídos como apoio para a extração de padrões na mineração de textos

Daniel Luiz de Albuquerque[1]
Solange Oliveira Rezende[1]
Roberto Hiroshi Higa[2]
Maria Fernanda Moura[2]

A quantidade de dados no mundo cresce imensamente a cada dia. A análise de grandes conjuntos de dados, também conhecida por *Big Data*, consiste em uma base fundamental para o aumento de produtividade e inovação (MANYIKA et al., 2011). Particularmente, a análise de dados textuais corresponde à mineração de textos, que tem por objetivo encontrar tendências, padrões e conhecimento em documentos textuais escritos em linguagem natural (EBECKEN et al., 2003). As etapas de um processo genérico de mineração de textos se dividem em: a) identificação do problema; b) pré-processamento; c) extração de padrões; d) pós-processamento; e) utilização do conhecimento.

Neste trabalho, o foco está na extração de padrões e avaliação objetiva (pós-processamento), no projeto "Tecnologias para computação distribuída, armazenamento de grandes volumes de dados e *workflow* científico, em suporte à pesquisa agropecuária", do Macroprograma 5 da Empresa Brasileira de Pesquisa Agropecuária (Embrapa), no seu plano de ação de "Ferramentas de mineração de dados aplicadas às áreas de bioinformática e mudanças climáticas" e atividade "Avaliação de ferramentas de mineração de textos para a arquitetura Hadoop/MapReduce". Dessa forma, o principal foco deste trabalho é a identificação, avaliação e configuração de um conjunto de ferramentas adequadas aos processos de mineração

[1] LABIC (Universidade de São Paulo) - daniel.albuquerque@colaborador.embrapa.br
[2] LABIC (Embrapa Informática Agropecuária) – {roberto.higa, maria-fernanda.moura}@embrapa.br

de textos em ambientes distribuídos para os domínios de bioinformática e mudanças climáticas.

O primeiro experimento conduzido para teste do uso do ambiente distribuído consiste em utilizar algoritmos de agrupamento para identificar grupos e subgrupos de genes. Considerando determinadas desordens relacionadas aos genes como classes, a avaliação do agrupamento hierárquico será capaz de identificar se as informações textuais são suficientes para identificar grupos de genes relacionados a uma desordem. Na identificação do problema, etapa na qual são definidos objetivos, métodos, base de dados e algoritmos a serem utilizados, também foi identificado outro problema, o tamanho da base de dados.

De acordo com o estudo realizado por Zhao e Karypis (2002) acerca da complexidade computacional dos algoritmos de agrupamento hierárquico, constatou-se que estes têm sua aplicação limitada pela quantidade de dados. Considerando que a grande maioria dos algoritmos estudados, tais como o *Bissecting-Kmeans* (divisivo) e o *Average-Linkage* (aglomerativo), possuem complexidade igual ou superior a O (*n log n*), torna-se inviável aplicar tais algoritmos de forma centralizada ao problema apresentado. Este problema motivou o estudo e aplicação do Framework de sistemas distribuídos denominado Hadoop e um projeto associado denominado Mahout, o qual consiste em uma biblioteca com diversas implementações de algoritmos de aprendizado de máquina utilizando-se um modelo de programação distribuída fornecido pelo Hadoop/MapReduce (YANG et al., 2007).

Considerando que o Mahout utiliza o Hadoop, o primeiro passo consiste em criar e configurar um *cluster*, para o processamento *multi-node*. Além do processamento ser paralelizado, os dados também devem estar armazenados em diferentes nós do cluster, utilizando o *Hadoop Distributed File System* (HDFS), o sistema de arquivos do Hadoop. Feita a configuração, deverão ser executados e avaliados os algoritmos de agrupamento implementados no Mahout. Para cada algoritmo deverão ser exploradas as diferentes configurações e observados os tempos de execução. Também será analisada a escalabilidade do processamento paralelo para diferentes números de máquinas, dado que esta característica foi a principal motivadora pela escolha desta metodologia. A Figura 1 ilustra a configuração *multi-node*, na qual existe um nó *master* para administrar o processamento e os dados armazenados em *N* nós *slaves*.

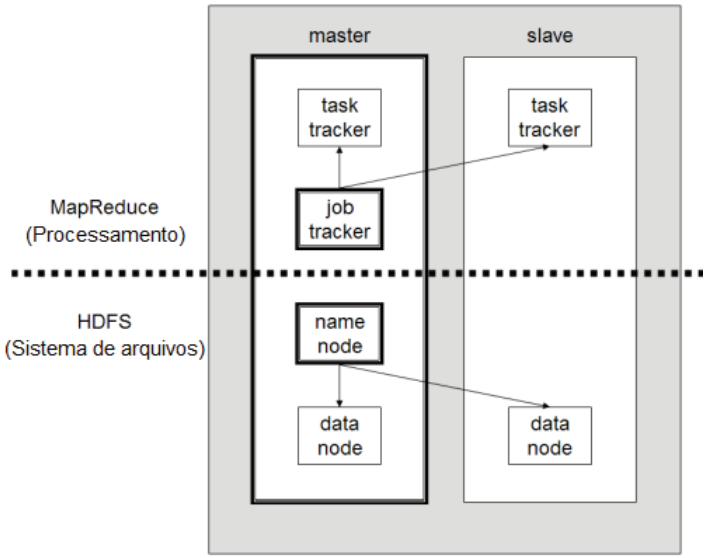

Figura 1. Configuração de um cluster no Hadoop.
Fonte: Noll (2013).

Para avaliar os agrupamentos gerados será utilizada a medida de avaliação F-Score, a qual utiliza o conceito de recuperação em agrupamentos hierárquicos de documentos através dos cálculos de precisão e revocação (ZHAO; KARYPIS, 2002).

Espera-se como resultados deste trabalho: um tutorial de instalação da arquitetura e ferramental, bem como configurações dos algoritmos de aprendizado de máquina de interesse do projeto; e, a avaliação do agrupamento hierárquico de textos para identificar grupos de genes relacionados a uma desordem.

Referências

EBECKEN, N. F. F.; LOPES, M. C. S.; ARAGÃO M. C. D. Mineração de textos. In: REZENDE, S. O. (Ed.). **Sistemas Inteligentes**: fundamentos e aplicações. Barueri: Manoele, 2003. p. 337–364.

MANYIKA, J.; CHUI, M.; BROWN, B.; BUGHIN, J.; DOBBS, R.; ROXBURGH, C.; BYERS, A. H. **Big data**: the next frontier for innovation, competition, and productivity. [S.l.]: McKinsey Global Institute, 2011.

NOLL, M. G. **Running Hadoop on Ubuntu Linux (multi-node cluster)**. 2013. Disponível em: <http://www.michael-noll.com/tutorials/running-hadoop-on-ubuntu-linux-multi-node-cluster/>. Acesso em: 27 set. 2013.

YANG, H. C.; DASDAN, A.; HSIAO, R. L.; PARKER, D. S. Map-reduce-merge: simplified relational data Processing on large clusters. In: ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 2007, Beijing. **Proceeding**... New York: ACM, 2007. p. 1029-1040.

ZHAO, Y.; KARYPIS, G. Evaluation of hierarchical clustering algorithms for document datasets. In: INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 11., McLean, VA. **Proceedings**... New York: ACM, 2002. p. 515-524.