# Mineração de Textos: Uma proposta de Workflow para Pré-processamento de Documentos em Grid Computacional

Emerson da Silva Borges[1]; Thiago Ferauche[2]; Profº Dr. Maurício Amaral de Almeida[3]

Centro Estadual de Educação Tecnológica Paula Souza (CEETEPS) - São Paulo – Brasil Julho de 2010

(1)borges.emerson@ymail.com; (2)thiago.ferauche@gmail.com; (3)madealmeida@gmail.com

**Resumo –** A mineração de textos é um processo que demanda recursos computacionais mais intensos do que a recuperação de informação de dados pré-estruturado e requer a transferência e filtragem de grande quantidade de dados. Os ambientes de Grid computacionais proveem uma infraestrutura adequada para acomodar essas tarefas. A computação em Grid consiste num conjunto de serviços, protocolos e *software* que integra recursos operacionais distribuídos de forma unificada. Este artigo apresenta uma proposta de *workflow* para otimização da etapa de pré-processamento de documentos da mineração de textos, tomando por base textos da Justiça Trabalhista, utilizando o ambiente de grid computacional e ferramentas de mineração de textos.

**Palavras chave**: Descoberta de conhecimento em bases textuais, Mineração de Textos, Grids Computacionais.

## Introdução

Estudos indicam que 80% dos documentos gerados pelas organizações no mundo é representada por documentos textuais. Neste cenário técnicas automatizadas e semi-automáticas de Mineração de Textos (MT) são aliados importantes para suporte a extração de conhecimento relevante com base nestes conteúdos. Serviços de MT é parte da visão estratégica do Knowlegde Grid [1]. A necessidade de soluções de mineração de texto tem sido reconhecida pelo Datamining Grid [ 11].

PRETEXT [6] é uma ferramenta computacional implementada em Perl [7], desenvolvida com o objetivo de realizar o pré-processamento de um conjunto de documentos utilizando a abordagem *bag-of-words.* Utilizaremos PRETEXT no nosso experimento devido a sua eficiência na redução da dimensionalidade e geração de tabela atributo-valor que serve de base para os algoritmos de mineração de dados.

Esse trabalho apresenta uma proposta de workflow de pré-processamento de documentos paralelizáveis em ambiente de Grid Computacional, utilizando o PRETEXT. A seção 2 descreve o ambiente, ferramentas e metodologia para execução do experimento. Após essa definição descrevemos a proposta do experimento. Fechamos o artigo com as referências bibliográficas do documento.

## 2. Ambiente, Ferramentas e Metodologia

Neste tópico descrevemos as características básicas do ambiente Grid que pretendemos utilizar em nosso experimento, seguido de uma descrição das ferramentas de MT que abordamos e a metodologia que estamos introduzindo em nosso TM workflow.

### 2.1. Ambiente de Grid Computacional

O Globus Toolkit [2,4] provê um caminho seguro e flexível para acesso seguro e distribuição dos recursos. Ele permite a usuários acessar sistemas computacionais com diferentes sistemas de hardware e software de maneira padronizada, ainda que estes sistemas estejam localizados em diferentes locais geográficos. Utilizaremos este m*iddleware* em nosso experimento. Para a transferência de arquivos entre os nós e clientes da rede do ambiente estamos utilizando o GridFTP [4] . No processo de gerenciamento dos *jobs,* integramos o globus ao PBS Professional [5], um gerenciador de jobs para ambientes paralelizáveis como *clusters e grids computacionais.* O PBS (Portable Batch System) permite a submissão e gerenciamento de jobs paralelizáveis em ambientes clusterizados com múltiplos processadores, como é o caso de Grids Computacionais [10].

### 2.2. Ambiente de Mineração de Textos

No nosso sistema de Grid fizemos a instalação da ferramenta Pretext [6] nos dois nós do ambiente. Essa configuração facilita a execução do nosso *workflow* de pré-processamento paralelizável de documentos. Estando a ferramenta e mineração de textos disponível nos dois servidores do Grid é possível executar várias instâncias de pré-processamento de textos com o PBS [5] devolvendo os arquivos de resultados pelo GridFTP a máquina cliente da solicitação do serviço TM. Há necessidade também da ferramenta, no nosso caso o Pretext, ser interpretada possibilitando que seja um *UNIX job.* A figura 1 detalha o fluxograma do processo de mineração de textos executando em múltiplos nós de Grid, gerenciados pelo GRAM integrado ao gerenciador de *jobs* PBS [5].

### 2.3. Metodologia

Implementado o ambiente de Grid e pré-processamento de textos propusemos o workflow de MT exbido na figura 1. Para gerenciamento dos recursos do Grid utilizamos o GRAM do globus junto com o PBS [5]. Recebemos os *corpus* de textos da Justiça Trabalhista, através do GridFTP [4] que são submetidos ao agendador de jobs para que sejam paralelizados e executados de acordo com a disponibilidade dos nós do ambiente. Todos os dados são transferidos através de uma área de transferência comum compartilhada por NFS (Network File System). A medida que vão sendo executados os *jobs*, que são instâncias de um processamento configurados de pré-processamento de documentos

pelo PRETEXT, cada nó dispõe de um serviço ( processo nó ) que monitora e envia os arquivos de resultado da TM ao módulo de controle. Todo o controle dos *corpus* de documentos textuais bem como os arquivos de resultados são monitorados pelo fluxo do Processo Controle. Uma visão macro do *workflow* da proposta pode ser visualizado na Figura abaixo

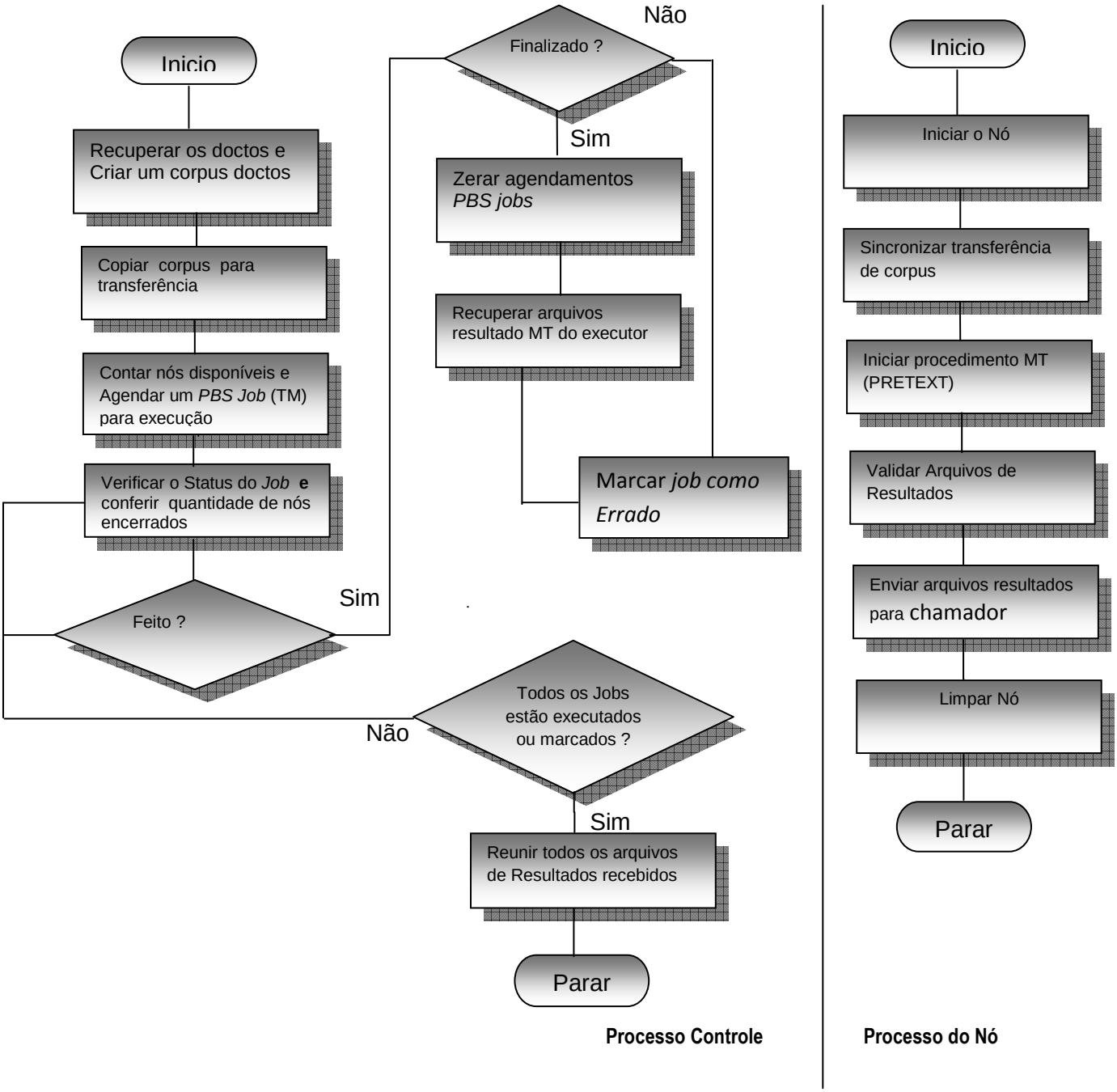


**Figura 1**. *Workflow* do processo de pré-processamento de MT, com PRETEXT em Grid Computacional

## 3. Considerações Finais

Nossa proposta é exequível, em ambiente computacional com a configuração mínima de dois servidores Linux, com processadores Pentium III de 800 Mhz, 1 Gigabyte de memória RAM, para o ambiente Grid. Utilizamos o *middleware* Globus Toolkit 4 [2]. Utilizaremos no experimento duas máquinas com o sistema operacional Debian [3] para o ambiente de Grid, ambas com PRETEXT instalado.

Os experimentos da aplicação dos *corpus* de documentos da Justiça Trabalhista no ambiente aqui descrito, serão realizados com o propósito de obter processos de mineração de textos paralelizáveis. Estes estudos serão desenvolvidos neste próximo ano.

## 4. Referências


[1] M. Cannataro and D. Talia. **The knowledge grid. Communications of the ACM**, 46/1, pages 89 – 93, 2003.http://grid.deis.unical.it/kgrid/.

[2] Globus Toolkit 4. *Web site :* <http://www.globus.org/toolkit/docs/4.0/ >

[3] Debian. *Web site* 16 de julho de 2010: <http://www.debian.org> .

[4] GridFTP. Web site. 20 de junho de 2010: <http://www.globus.org/toolkit/docs/4.0/data/gridftp/ >.

[5] PBS Pro. Web site. 26 de junho de 2010: <http://www.pbsworks.com/ >.

[6] PRETEXT. *Web site* 16 de junho de 2010: <http://www.labic.icmc.usp.br/pretext2/ >.

[7] Perl. *Web site 28 de julho de 2010:* <http://www.perl.org/ >.

[8] FELDMAN, R. & SANGER, J. **The Text Mining HandBook – Advanced Approaches in Analyzing Unstrutected Data**, Editora Cambridge, Cambridge, 2007

[9] I. FOSTER and C. KESSELMAN. **Globus: A metacomputing infrastructure toolkit**. International Journal of Supercomputer Aplications 11, 2, 115-128. *FTP site 20-07-2010*:

<ftp://ftp.globus.org/pub/globus/papers/globus.pdf >

[10] I. FOSTER, C. KESSELMAN, and S. TUECKE. **The anatomy of the grid: Enabling scalable virtual organizations**. *Web site 20-07-2010*:

< http://www.globus.org/alliance/publications/papers/anatomy.pdf >

[11] DataMiningGrid. *Web site*. 29 jun 2010: <http://www.datamininggrid.org/ >