Mineração de Textos: Pré-processamento Distribuído de Documentos para Algoritmos de Aprendizagem de Máquina

Emerson da Silva Borges

Centro Estadual de Educação Tecnológica Paula Souza - São Paulo – Brasil

borges_emerson@yahoo.com.br

Profº Dr. Maurício Amaral de Almeida

Centro Estadual de Educação Tecnológica Paula Souza - São Paulo – Brasil

madealmeida@gmail.com

Abstract – This paper presents a methodology to optimize the preprocessing step for text mining of documents, based on texts of Justice Labor, using the grid computing environment or clusters dedicated or opportunistic and text mining tools.

Keywords: Knowledge Discover in Text Bases, Distributed Computing, Griding Computing, Text Mining

Resumo – Este artigo apresenta uma proposta metodológica de 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, aglomerados dedicados ou oportunistas e ferramentas de mineração de textos.

Palavras chave: Descoberta de conhecimento em bases textuais, Computação distribuída, Grids Computacionais, Mineração de Textos

Introdução

Estudos indicam que 80% dos documentos gerados pelas organizações no mundo é representada por documentos textuais [1]. 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.

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 provem uma infra-estrutura 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 [2].

Este artigo apresenta uma proposta metodológica de pré-processamento de documentos paralelizáveis em ambiente de Grid Computacional [3] ou aglomerados computacionais, utilizando a técnica de escalonamento de tarefas. Foram utilizados técnicas aplicáveis tanto a Grids Computacionais Oportunistas quanto a ambientes de Grids Computacionais dedicados [3]. O primeiro caso de Grid que trabalha com o reaproveitamento de recursos ociosos conectados pela infra-estrutura de Grids disponível.

Foram utilizadas técnicas de escalonamento global de tarefas [4]. A pesquisa adota também o emprego de computação paralela, uma técnica que normalmente é aplicada em computação distribuída, que consiste em quebrar um problema numa coleção de pequenas tarefas e, então submetê-las para que sejam executadas em diferentes computadores [5].

Trabalhos de pesquisa semelhantes a este tem estudado a otimização de etapas do processo de mineração de textos (MT) com a utilização de computação distribuída em Grids Computacionais Dedicados ou oportunistas e também utilizando os aglomerados computacionais dedicados ou oportunistas [6,7] . O projeto de pesquisa do Grid Unicore europeu ligado ao Ministério da Educação da Alemanha desenvolveu uma pesquisa de mineração de textos multi-cluster [6] onde o foco é na utilização de nós do Grid D-Core para análise de textos biomédicos com base nos dados da PubMed[1]. Outra pesquisa interessante nesta área é a realizada no COPPE da Universidade Federal do Rio de Janeiro (UFRJ) [7] que consiste na classificação de documentos utilizando Grids Computacionais.

Metodologia

O procedimento metodológico toma por base as quatro etapas do processo de mineração de textos [8]. Como o foco da pesquisa é o pré-processamento de documentos em computação distribuída, com a coleção de documentos do trabalho de Ferauche [9] utiliza a primeira e segunda etapa desta metodologia:

[1] PubMed é uma base de conteúdo da literatura biomédica do *National Center for Biotechnology Information* , EUA *National Library of Medicine* composto por mais de 21 milhões de citações para a literatura biomédica em MEDLINE, revistas de ciências da vida, e livros online.

1. **Coleta de Documentos**: nesta fase os documentos relacionados com o domínio da aplicação final são selecionados.
2. **Pré-processamento:** consiste de um conjunto e ações realizadas sobre o conjunto de textos obtidos na etapa anterior, com o objetivo de prepará-los para a extração do conhecimento.
3. **Extração de Conhecimento:** utilizam-se alguns algoritmos de aprendizado com o objetivo de extrair, a partir de documentos pré-processados, conhecimento na forma de regras de associação, relações, segmentação, classificação de textos, entre outros.
4. **Avaliação e Interpretação dos Resultados:** nessa etapa os resultados obtidos são analisados, filtrados e selecionados para que o usuário possa ter um melhor entendimento dos textos coletados. Esse entendimento maior pode auxiliar em algum processo de tomada de decisão.

Este trabalho realiza a primeira e segunda etapas do processo de mineração de textos. Ao realizar o procedimento de pré-processamento (2ª etapa) é realizado o processamento em uma máquina local e, então repetido agora, utilizando técnicas de computação distribuída, buscando otimizar o desempenho computacional da tarefa

Coleta de Documentos

Esta etapa é realizada utilizando uma coleção de documentos que é composta por textos extraídos das ementas processuais do Tribunal Regional do Trabalho da 2ª Região - SP. Uma ementa é um resumo de uma decisão (Acórdão) tomada por um colegiado de desembargadores federais. A coleção de ementas utilizada foi disponibilizada pela pesquisa de Ferauche. Os arquivos utilizados para o processo de pré-processamento deste trabalho correspondem aos meses de janeiro a dezembro, dos anos de 2008 a 2010. Estes textos são divididos em 187 categorias de decisões e alguns exemplos são visualizados na tabela1. Nesta etapa eles não seguem uma distribuição equânime. Para a seleção das categorias de documentos que foram utilizados na etapa de pré-processamento, respeitou-se o no a distribuição dos documentos em categorias a teoria de PAC-Learning [10], que consiste que para termos um bom aprendizado há necessidade de uma boa distribuição dos exemplos de treinamento.

Tabela 1: Exemplos de categorias de ementas trabalhistas

CATEGORIA	QUANTIDADE DE DOCUMENTOS	TAMANHO (Bytes)
PREVIDENCIA SOCIAL	12865	10.250.190
EXECUÇÃO	5370	4.191.805
MÃO-DE-OBRA	4308	4.743.027
EMBARGOS DECLARATÓRIOS	4248	2.561.049
PROVA	3689	2.867.747
RELAÇÃO DE EMPREGO	2922	2.583.514
PRESCRIÇÃO	2834	2.728.237
DANO MORAL E MATERIAL	2532	2.529.515
COMPETÊNCIA	2151	2.397.916
SINDICATO OU FEDERAÇÃO	2094	2.079.449

Pré-processamento

Nesta etapa é necessário a preparação dos documentos, para obter um conjunto de características dos mesmo através da abordagem estatística, que formam um vetor atributo-valor, onde cada termo é um atributo do vetor. Para a realização deste procedimento foi utilizado o PRETEXT II [11] que é uma ferramenta computacional implementada do LABIC-USP, desenvolvida com o objetivo de realizar o pré-processamento de um conjunto de documentos não estruturados utilizando a abordagem bag-of-words.

É utilizado PRETEXT II no experimento devido a sua eficiência na redução da dimensionalidade e geração de tabela atributo-valor que serve de base para os algoritmos de aprendizado de máquina utilizados na etapa 3 da mineração de textos.

Tabela 2. Distribuição das Categorias para pré-processamento distribuído

Classe	Tot. Doc	Usados	Outros	Total Docs	Usados
EXECUÇÃO	5370	**3000**	EMBARGOS DECLARATÓRIOS	4248	1083
			RELAÇÃO DE EMPREGO	2922	745
			SINDICATO OU FEDERAÇÃO	2094	534
			MANDADO DE SEGURANÇA	1612	411
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	229
			Total de Outros	11771	3002
MÃO-DE-OBRA	4308	**3000**	PROVA	3689	1272
			SINDICATO OU FEDERAÇÃO	2094	722
			RECURSO	1297	448
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	309
			PROCESSO	726	251
			Total de Outros	8701	3002
PREVIDENCIA SOCIAL	12865	**3000**	EMBARGOS DECLARATÓRIOS	4248	983
			DANO MORAL E MATERIAL	2532	586
			PROVA	3689	853
			MANDADO DE SEGURANÇA	1612	373
			RESPONSABILIDADE SOLIDÁRIA/SUBSIDIÁRIA	895	207
			Total de Outros	12976	3002
EMBARGOS	4248	3000	PRESCRIÇÃO	2834	1022
DECLARATÓRIOS			SINDICATO OU FEDERAÇÃO	2094	755
			CONCILIAÇÃO	1377	497
			HORAS EXTRAS	1136	410
			NORMA COLETIVA (EM GERAL)	885	319
			Total de Outros	8326	3003
PROVA	3689	3000	COMPETÊNCIA	2151	1095
			ASSITÊNCIA JUDICIÁRIA	1564	796
			HORAS EXTRAS	1136	579
			TEMPO DE SERVIÇO	592	302
			RESPONSABILIDADE	452	231
			Total de Outros	5895	3003

Os textos foram pré-processados em duas etapas:

1. Nesta primeira etapa os textos foram pré-processados numa máquina Intel Pentium de 3 GHz, com dois núcleos. Nesta máquina foi instalado um gerenciador de máquina virtual para os testes. Foi utilizada a ferramenta PRETEXT II localmente. Cada uma das categorias da tabela 1 foram pré-processadas por vez.
2. Na segunda etapa, todas as categorias foram submetidas a uma fila de execução com a ferramenta de escalonamento de tarefas Condor [12], aplicada para o gerenciamento de Grids oportunistas. Essa solução permite a execução do PRETEXT II utilizando a técnica de paralelismo descrita por Wilkinson, que faz a proposta de dividir o problema em partes menores para execução distribuída.

Foram utilizadas duas máquinas com a configuração igual a máquina utilizada na primeira etapa. O Condor[2] é um sistema de gestão especializado na carga de trabalho empregado na computação distribuída de uso intensivo. Como outros sistemas com características em lote, ele fornece um mecanismo de fila de trabalho, política de programação, esquema de prioridade, monitoramento de recursos e gerenciamento de recursos.

Análise e Resultados

Os seguintes resultados foram nas tabelas 3 e 4 da pesquisa :

Tabela 3: Dados de processamento do teste 1

Categoria	Qtde Doctos[3]	Outros[4]	Total de doctos por categoria	Qtde de Nodes [5]	Tempo[6]
MAO DE OBRA	3000	3002	6002	1	741
EMBARGOS	3000	3003	6003	1	500
EXECUCAO	3000	3002	6002	1	589
PREVIDENCIA	3000	3002	6002	1	479
PROVA	3000	3003	6003	1	574
	15000	15012	30012		**2883**

Tabela 4: Dados de processamento do Teste 2

Classe	Documentos utilizados	Outros	Total de documentos	Qtd. Nodes	Tempo paralelizado (s)
EMBARGOS	3000	3003	6003	2	426,0207612
EXECUÇÃO	3000	3002	6002	2	357,5086505
MAO DE OBRA	3000	3002	6002	2	434,1176471
PREVIDENCIA	3000	3002	6002	2	365,6055363
PROVA	3000	3003	6003	2	342,5605536
	15000	15012	30012		**1800**

A coleção de documentos utilizada para a execução dos testes é a mesma, portanto tanto a distribuição das cinco categorias selecionadas da tabela 3, quanto as quantidades de documentos selecionados para os dois testes seguem a teoria de distribuição de treinamento de PAC-Learning, e também houve uma

[2] Escalonador de tarefas para computação distribuída de uso intensivo disponível em http://www.cs.wisc.edu/condor
[3] Quantidade de documentos utilizados da categoria selecionada. Exemplo: 3000 documentos da categoria Mão de Obra
[4] Quantidade de documentos selecionados de categorias diferentes da categoria em questão em quantidade proporcional para posterior treinamento
[5] O termo Node é utilizado para nomear uma unidade de máquina multiprocessada em computação distribuída
[6] Os tempos de processamento neste experimento foram medidos em segundos

atenção com a equivalência de documentos para igualar o volume de documentos para processamento nos dois testes.

Em primeira análise observa-se que houve uma redução nos tempos de processamento, tanto das estimativas por categoria, quanto no tempo total de processamento. Observa-se ainda que o tempo de processamento em duas máquinas separadas apresentaram um tempo superior ao tempo das duas máquinas utilizadas em paralelo.

Conclusão

Os experimentos da aplicação da coleção de documentos da Justiça Trabalhista no ambiente aqui descrito, estão sendo realizados com o propósito de obter processos de pré-processamento de textos paralelizáveis. Estes estudos estão em andamento, portanto foi apresentado neste artigo os resultados iniciais da pesquisa.

Nos resultados foram verificados uma tendência de ao processar todas as categorias juntas em paralelo, haver um gasto computacional (teste 2) em torno de 60% (sessenta por cento) do tempo computacional de processá-las separadamente em uma máquina (teste 1). A pesquisa prosseguirá e, até o momento há a justificativa de não trabalhar com Grids Computacionais Dedicados, comparando ao custo dos Grids Computacionais Oportunistas. Apesar do desempenho inferior está existe uma economia financeira quando trabalha-se com computação ambiente de computação distribuída oportunistas, porque utilizando este modelo não há necessidade de investimento na compra de novos computadores, considerando que esta abordagem computacional reutiliza processamento de máquinas ociosas no ambiente de Grid Computacional.

Os próximos trabalhos aumentarão a quantidade da coleção de textos e, também utilizará nos experimentos um número maior de nodes computacionais para avaliar o desempenho computacional da presente proposta.

Referências

[1] Tan, A.H (1999). Text mining: The state of art and the challenges. In Proceddings, PAKDD´99 workshop on Knowledge Discovery from Advanced Databases, Beijing, pp. 65-70.

[2] Dantas, M. Computação Distribuída de Alto Desempenho. Editora Axcel: Rio de Janeiro 2005.

[3] Foster, I.; Kesselman, C.; Nick, J. M.; Tuecke, S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations. The International Journal of High Performance Computing Applications, Disponível em <

http://www.globus.org/alliance/publications/papers/anatomy.pdf>. Acessado em site 20-07-2010.

[4] Casavant, T. L.; Kuhl, J., G. (1988) A Taxonomy of Sheduling in General-Purpose Distributed Computing Systems. IEEE Transaction on Software Engineering, v.14, n.2, p. 141-154, Feb 1988.

[5] Wilkinson, B. (2010) Grid Computing: Techniques and Applications. University of North Carolina, Charlotte, USA: 2010. 1 ISBN: 9781420069532.

[6] Kumpf, K., Mevissen, T., Waeldrich, O., Ziegler, Wolfgang (2007), Multi Cluster Text Mining on the Grid using D-Grid UNICORE environment, CoreGRID TR-0109

[7] Roncero, V.G., Costa, M.C.A., Ebecken, N.F.F. (2010), Text Classification on a Grid Enviroment, Centro de Tecnologia – UFRJ - Rio de Janeiro – RJ - Brasil

[8] Ferauche, T. (2011) Aprendizado de Classificadores de ementas da Jurisprudência do Tribunal Regional do Trabalho da 2ª Região – SP. VI WorkShop de Pesquisa do Centro Estadual de Eucação Tecnológica Paula Souza – SP – Brasil.

[9] Martins, C.A. (2003), Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado, Tese de doutorado – ICMC-USP

[10] Russel, S.; Norvig, P. (2004). Inteligência Artificial: tradução da 2ª ed. Rio de Janeiro: Elsevier.

[11] Soares, M.V.B, Prati, R.C., Monard, M.C. (2008). PRETEXT II: Descrição da Reestruturação da Ferramenta de Pré-processamento de Textos, ICMC-USP, São Carlos – SP – Brasil.

[12] Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny, "Condor - A Distributed Job Scheduler", in Thomas Sterling, editor, Beowulf Cluster Computing with Linux, The MIT Press, 2002. ISBN: 0-262-69274-0.

Contato

Emerson Borges, pesquisador de Computação Disribuída de Alto Desempenho e Suporte Técnico do Ministério Publico Federal da 3ª Região

E-mail: borges_emerson@yahoo.com.br