Framework de mineração de textos para análise de dados de bioinformática

Rafael Seiji Tanaka[1]
Rafael Hideo Kashiwara[1]
Roberto Hiroshi Higa[2]
Maria Fernanda Moura[2]

O projeto Prosgen (código SEG 03.09.01.0.25.00.00) se propõe a utilizar informações oriundas do banco de dados de artigos científicos Pubmed (PUBMED, 2011) e técnicas de mineração de textos para apoiar a interpretação biológica de genes candidatos. Seu objetivo é construir um framework de mineração de dados que suporte os processos de priorização e prospecção de genes candidatos associados a características de interesse econômico para a agricultura, para posterior introdução em programas de melhoramento genético. Neste trabalho, apresenta-se o estágio atual de desenvolvimento deste framework, com foco na interpretação de dados de expressão gênica por microarranjos.

Para construção do framework, diferentes programas e bibliotecas de software estão sendo desenvolvidos e/ou integrados (Figura 1). O programa eutils-search foi implementado com a finalidade de fazer o download do site do Pubmed de uma grande quantidade de artigos, relacionados a um organismo específico ou não, e armazená-la em um banco de dados local. Para o seu funcionamento, são necessários 4 arquivos que podem ser obtidos no site do NCBI (NCBI, 2011): (i) o arquivo gene2pubmed, que relaciona os genes de um organismo com um conjunto de publicações do Pubmed ; (ii) o arquivo gene_info, que contém anotações básicas sobre genes como seu ID, descrição e símbolo; (iii) o arquivo generifs_basic, que contém informação sumarizada, fornecida por usuários, sobre a função de genes; e (iv) o

[1] *Universidade Estadual de Campinas, rafaelst@cnptia.embrapa.br, rafakashiwara@gmail.com*

[2] *Embrapa Informática Agropecuária, {roberto, fernanda}@cnptia.embrapa.br*

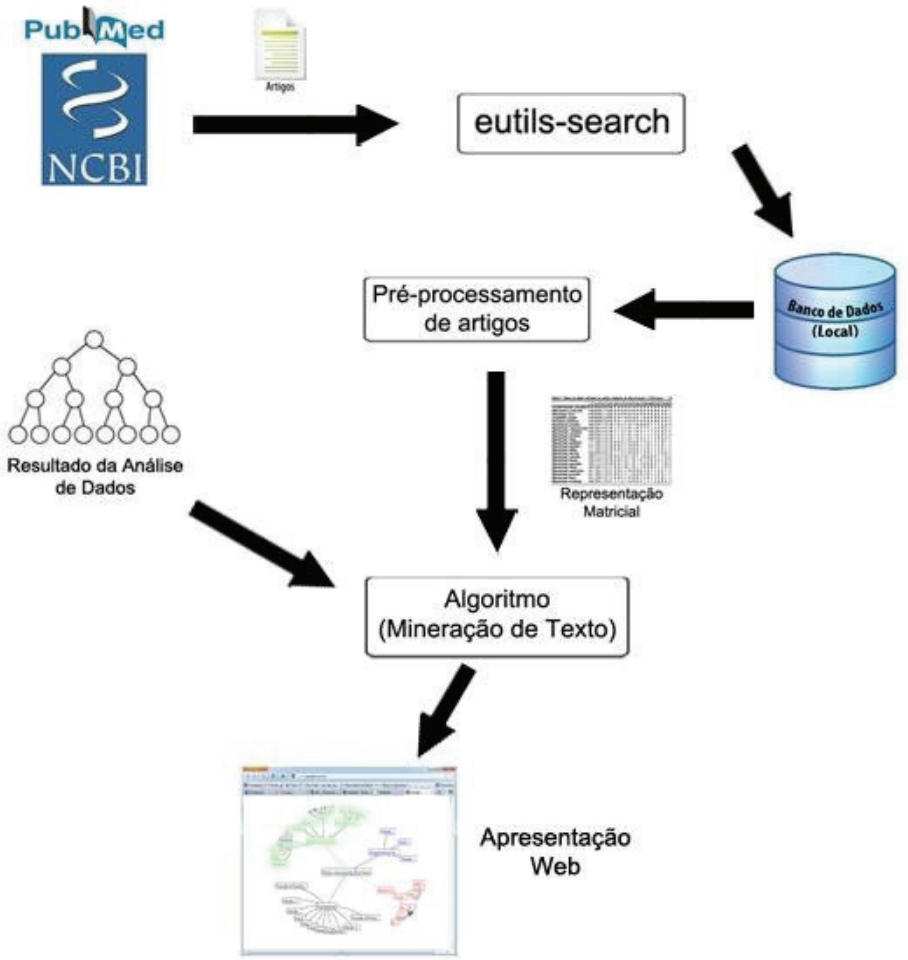

Figura 1. Processo de análise de dados utilizando mineração de textos.

arquivo nodes.dmp, que especifica as relações taxonômicas entre os diversos organismos. Para utilizar o eutils-search, o usuário informa o organismo a que os artigos estão relacionados. O programa, então, verifica quais são os genes relacionados, identifica os correspondentes artigos e faz seu download para a base de dados local PostgreSQL (POSTGRESQL, 2011). Esse procedimento é relativamente demorado, devido às restrições impostas pelo PubMed para acesso ao seu banco de dados.

A partir do banco de dados local, gera-se um conjunto de documentos em formato XML que é submetido a uma série de procedimentos de pré-processamento. Para realizar esse processo foi construída uma biblioteca java, baseada no projeto Apache Lucene (LUCENE, 2011), compreendendo as seguintes funcionalidades: (i) parseamento dos arquivos; (ii) construção de *stems*; (iii) remoção de *stopwords*; (iv) construção e seleção de n-gramas e (v) contagem de frequências de ocorrência. Ao final da fase de pré-processamento, obtém-se uma representação vetorial dos documentos XML, por meio de uma matriz Termo x Documento.

A partir deste ponto, diferentes processamentos são possíveis. O trabalho atualmente em desenvolvimento considera o problema de análise de cluster de dados de expressão gênica por microarranjos. Considera-se uma hierarquia de genes, onde um ou mais artigos pré-processados podem estar associados a cada gene. Está sendo desenvolvido um algoritmo para rotular os ramos da hierarquia com base no conteúdo da matriz termo x documento, isto é, encontrar descritores adequados a cada grupo.

A saída desse processo é um arquivo XML que representa a hierarquia de genes com seus ramos rotulados com os termos estatisticamente associados ao conjunto de genes associados. Para visualização desse resultado, está sendo desenvolvida uma interface web, baseada em um *applet* de árvore hiperbólica (Treebolic) (BOU, 2011).

A utilização eficiente da literatura científica disponível é essencial na pesquisa biológica, tanto na fase de planejamento experimental quanto na interpretação dos resultados obtidos. Em particular, na análise de dados de bioinformática, em que muitas entidades (ex: genes) são analisadas simultaneamente, a agregação automática da literatura ao processo, com uma breve sumarização desta (automaticamente obtida), pretende auxiliar qualitativamente a identificação das funções biológicas compartilhadas por grupos de genes e a comparação entre genes, baseado em sua descrição funcional.

Referências

BOU, B. Treebolic 2. 2011. <http://treebolic.sourceforge.net/en/index.html>. Acesso em: 1 out. 2011.

LUCENE. 2011. Disponível em: <http://lucene.apache.org/>. Acesso em: 1 out. 2011.

NCBI. National Center for Biotechnology Information. 2011. Disponível em: <http://www.ncbi.nlm.nih.gov/>. Acessado em: 1 out. 2011.

POSTGRESQL. Disponível em: <http://www.postgresql.org/>. Acesso em: 1 out. 2011.

PUBMED. Disponível em: <http://www.ncbi.nlm.nih.gov/pubmed>. Acesso em: 1 out. 2011.