Uma ferramenta para expansão do vocabulário com base em coocorrência

Exupério Lédo Silva Júnior[1]
Roberta Akemi Sinoara[2]
Solange Oliveira Rezende[3]
Ricardo Marcondes Marcacini[4]
Maria Fernanda Moura[5]

Resumo: Neste trabalho é apresentado um módulo desenvolvido para a experimentação de algumas técnicas de pré-processamento visando uma boa representação de coleções de documentos. As técnicas experimentadas são voltadas à expansão de vocabulário do domínio por meio da inclusão de termos coocorrentes. Um módulo, chamado DATool, foi desenvolvido em Java e experimentos estão sendo realizados. Caso os resultados sejam positivos, as técnicas serão transformadas em filtros de pré-processamento e indexação a serem incorporados ao arcabouço de ferramentas do projeto CRITIC@.

Palavras-chave: mineração de textos, pré-processamento de textos, coocorrência de termos.

[1] Estudante de Ciências de Computação da Universidade de São Paulo (ICMC-USP), estagiário da Embrapa Informática Agropecuária, Campinas, SP.

[2] Bacharel em Informática, doutoranda em Ciências da Computação e Matemática Computacional no Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (ICMC-USP), São Carlos, SP.

[3] Graduada em Licenciatura em Ciências Habilitação Matemática, doutora em Engenharia Mecânica, professora associada da Universidade de São Paulo (ICMC-USP), São Carlos, SP.

[4] Bacharel em Informática, doutor em Ciências da Computação e Matemática Computacional, docente e pesquisador da Universidade Federal de Mato Grosso do Sul (UFMS), Três Lagoas, MS.

[5] Estatística, doutora em Ciências Matemáticas e da Computação, pesquisadora da Embrapa Informática Agropecuária, Campinas, SP.

Introdução

Técnicas de Mineração de Textos auxiliam a extrair e organizar conhecimento de grandes coleções de documentos textuais, uma atividade cada vez mais importante dada a quantidade crescente de documentos textuais no meio digital (ROSSI, 2011). Tais técnicas baseiam-se na busca de padrões, tendências e regularidades em documentos escritos em língua natural.

O processo de Mineração de Textos pode ser dividido em cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento (REZENDE et al., 2003). Na identificação do problema são definidos os objetivos do processo de Mineração de Textos. Uma vez definidos o escopo do problema e o objetivo da aplicação, e selecionados documentos que representem o domínio do problema, partimos para a próxima etapa, o pré-processamento. Na etapa de pré-processamento, é realizado um tratamento dos dados. Tal tratamento engloba atividades como: eliminação de termos sem valor terminológico para o domínio em questão (as *stopwords*), normalização das palavras, identificação dos termos do domínio e seleção de atributos (LAGUNA, 2015; MOURA, 2009). Outra atividade importante realizada na etapa de pré-processamento é a estruturação dos dados, que devem ser representados em um formato apropriado para a extração de conhecimento. As tarefas a serem realizadas em seguida, na extração de padrões, consistem na utilização de algoritmos de aprendizado de máquina, que podem rotular, categorizar, ou detectar comportamentos intrínsecos da coleção de dados. O pós-processamento, por sua vez, consiste na avaliação dos padrões extraídos, verificando sua validade e aplicabilidade para que, no fim, o conhecimento extraído possa ser utilizado na etapa de utilização do conhecimento. Caso as atividades realizadas no pós-processamento indiquem que o conhecimento não atinge os objetivos estabelecidos, um novo ciclo do processo de Mineração de Textos é iniciado, com uma nova execução das etapas anteriores.

As atividades realizadas no pré-processamento são essenciais para o sucesso da Mineração de Textos, visto que a representação dos textos obtida nesta etapa deve manter os padrões a serem descobertos na extração de padrões. O módulo abordado neste artigo, chamado de DATool, é uma experimentação de técnicas para auxiliar as atividades de pré-processamento e indexação incremental de bases de dados textuais. Com a aplicação dessas técnicas, espera-se obter um conjunto maior de termos relacionados

ao domínio em questão e melhorar os resultados da Mineração de Textos. Na próxima seção são descritos os requisitos e funcionalidades do módulo desenvolvido. Logo após são apresentados os resultados obtidos até o momento e, por fim, as considerações finais.

Materiais e Métodos

O módulo DATool foi desenvolvido com o objetivo de permitir a experimentação de técnicas de pré-processamento e indexação de bases textuais no formato XML no contexto do projeto CRITIC@, que visa melhorar a gestão do conhecimento técnico-científico na área de recursos hídricos, por meio de análises cruzadas das informações, bem como subsidiar ações de investigação e disseminação do conhecimento na rede de pesquisa (SILVA; MOURA, 2014). Para o desenvolvimento desse módulo foram realizadas atividades de especificação dos requisitos, estudo das técnicas envolvidas (tanto do processo de Mineração de Textos, quanto de programação), implementação do módulo, documentação e testes. O módulo foi desenvolvido com base na ferramenta Torch (MARCACINI; REZENDE, 2010), permitindo o processamento incremental e a sua integração com outros métodos de Mineração de Textos.

Na Figura 1 são apresentados os tratamentos realizados nos textos de entrada para gerar a representação desses textos com base em um vocabulário controlado e termos coocorrentes, que pode ser incrementalmente atualizada a medida em que novos textos são adicionados à coleção.

O módulo recebe como entrada um arquivo de configuração fornecido pelo usuário que contém os parâmetros para execução do módulo. Nesse arquivo o usuário define: a) idioma da coleção de textos: sendo que o módulo suporta os idiomas português, inglês e espanhol; b) diretório de entrada: endereço do diretório contendo a coleção de textos; c) diretório de saída: endereço do diretório no qual são salvos os arquivos de saída; d) vocabulário inicial: endereço do vocabulário contendo as palavras que formam o vocabulário inicial; e) *stopwords*: endereço do arquivo contendo a lista de *stopwords*; f) *stemmização*: define se os termos são stemmizados ou não (*true* ou *false*); g) *tags*: nome das *tags* cujo conteúdo são considerados no pré-processamento; h) coocorrência: define os parâmetros para o cálculo dos coocorrentes: método e coeficiente de corte.

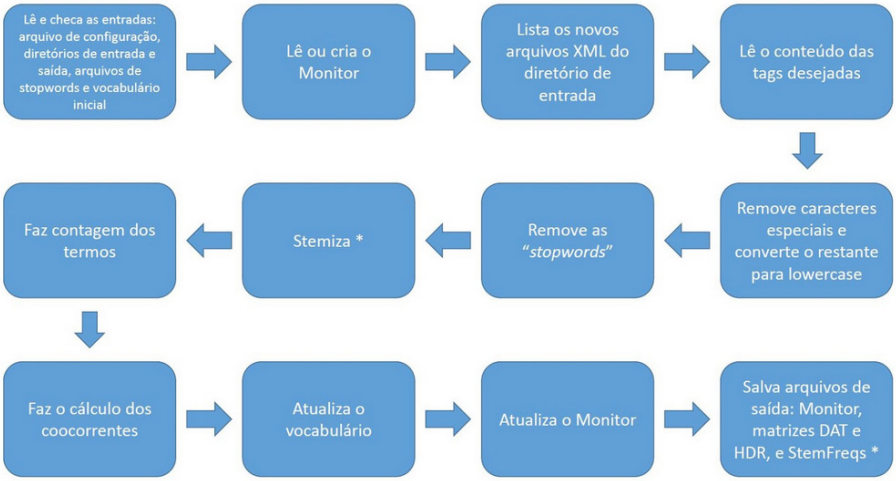

Figura 1. Fluxograma do módulo DATool.

Inicialmente o módulo lê e verifica todos os parâmetros passados pelo arquivo de configuração. Para garantir o aspecto incremental do módulo, utiliza-se um arquivo que guarda as informações dos processamentos anteriores, o Monitor. Então, caso seja a primeira execução, cria-se um novo Monitor; caso contrário, lê-se o Monitor existente. Os novos arquivos são listados e, para fins de normalização, o módulo realiza uma limpeza dos dados, selecionando apenas o conteúdo das tags desejadas, removendo os caracteres especiais e as stopwords, e facultativamente realizando a stemmização das palavras. A stemmização conta com o apoio da API Lucene[6] 5.2.0 da Apache, além da ferramenta Torch (MARCACINI; REZENDE, 2010). E o cálculo da coocorrência utiliza a API Commons Math[7] 3.5, também da Apache.

Posteriormente realiza-se a contagem dos termos por arquivo. Aqueles que também aparecem no arquivo de Vocabulário Inicial entram para o vocabulário. E então são calculados os coocorrentes por meio do coeficiente de correlação de Pearson. No caso da DATool, este coeficiente mede o grau de relação entre pares de termos do vocabulário e termos que estão fora do vocabulário. O cálculo leva em consideração a frequência destes termos

[6] Disponível em: <https://lucene.apache.org/>. Acesso em: 28 set. 2015.
[7] Disponível em: <http://commons.apache.org/>. Acesso em: 28 set. 2015.

em cada texto da coleção, e o resultado para o coeficiente varia entre [-1, 1]. Os termos que formam pares com o coeficiente maior, em módulo, que o coeficiente de corte definido no arquivo de configuração também são adicionados ao vocabulário.

No caso deste módulo, a saída é uma matriz atributo-valor, cujas linhas correspondem a cada documento da coleção de textos e as colunas a cada atributo selecionado, respeitando os padrões utilizados no projeto CRITIC@. Tal padrão consiste no arquivo texto HDR, que contém a lista dos documentos e a lista dos atributos, e o arquivo texto DAT, que informa a frequência dos atributos nos arquivos.

Após o desenvolvimento, foram realizados testes de unidade para validar as funcionalidades e os parâmetros do módulo. A ferramenta foi executada com uma coleção de 956 documentos em português no formato XML. A funcionalidade de *stemmização* também foi testada individualmente com outros textos escritos em espanhol e inglês.

Resultados e Discussão

Durante os testes de unidade, foi verificado que as funcionalidades implementadas estão funcionando conforme esperado, porém não foi feita uma análise sobre os resultados obtidos pela expansão do vocabulário. Atualmente, o módulo está em fase de teste e validação desses resultados. Os resultados esperados são que, dado um conjunto de termos inicial, outros termos também relacionados ao domínio sejam encontrados. Desta forma, é possível um grupo maior de termos relacionados ao domínio em questão, aumentando a representatividade da coleção de textos por meio do vocabulário.

Considerações Finais

Neste trabalho é apresentado o módulo DATool, que é uma experimentação de técnicas para expansão de vocabulário com base em coocorrência de termos para representação incremental de coleções textuais. Caso os experimentos que estão sendo realizados no momento tenham resultados

positivos, as técnicas validadas serão transformadas em filtros de pré-processamento e indexação a serem incorporados ao sistema do projeto CRITIC@.

Referências

LAGUNA, M. da S. C. **Extração automática de termos simples baseada em aprendizado de máquina**. 2014. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-11082014-103430/pt-br.php>. Acesso em: 28 set. 2015.

MARCACINI, R. M.; REZENDE, S. O. Torch: a tool for building topic hierarchies from growing text collections. In: WORKSHOP ON TOOLS AND APPLICATIONS, 9.; BRAZILIAN SYMPOSIUM ON MULTIMEDIA AND THE WEB, 2010, Belo Horizonte. **Proceedings**… [S.l.: s.n.], 2010. p. 1-3. WFA'2010; Webmedia 2010.

MOURA, M. F. **Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico**. 2009. Tese (Doutorado) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, SP. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/17875/1/MFM_Tese_5318963-2.pdf>. Acesso em: 28 set. 2015.

REZENDE, S. O.; PUGLIESI, J. B.; MELANDA, E. A.; de PAULA, M. F. Mineracão de dados. In: REZENDE, S. O. (Ed.). **Sistemas inteligentes**: fundamentos e aplicacões. Manole, 2003. p. 307-335.

ROSSI, R. G. **Representação de coleções textuais por meio de regras de associação**. 2011. Dissertação (Mestrado) - ICMC - USP - São Carlos. Disponível em: <http://www.teses.usp.br/teses/disponiveis/55/55134/tde-31082011-125648/pt-br.php>. Acesso em: 28 set. 2015.

SILVA, L. E. A.; MOURA, M. F. Componentes para a integração e extração de padrões em textos para versão 1.0 do ambiente CRITIC@. In: MOSTRA DE ESTAGIÁRIOS E BOLSISTAS DA EMBRAPA INFORMÁTICA AGROPECUÁRIA, 10., 2014, Campinas. **Resumos**... Brasília, DF: Embrapa, 2014. p. 17-19. Disponível em: <http://ainfo.cnptia.embrapa.br/digital/bitstream/item/118438/1/043-14.pdf>. Acesso em: 28 set. 2015.