Metodologia de Pré-processamento Textual para Extração de Informação em Artigos Científicos do Domínio Biomédico

Pablo Freire Matos[12]
Orientador: Ricardo Rodrigues Ciferri[2]
Coorientador: Thiago Alexandre Salgueiro Pardo[3]
{pablo_matos, ricardo}@dc.ufscar.br, taspardo@icmc.usp.br

[1]Programa de Pós-Graduação em Ciência da Computação

[1,2]Universidade Federal de São Carlos
Departamento de Computação – São Carlos, SP, Brasil

[3]Universidade de São Paulo
Departamento de Ciências de Computação – São Carlos, SP, Brasil

Nível: Mestrado
Ingresso: Março de 2008
Previsão de conclusão: Abril de 2010
Etapas concluídas: defesa da proposta em Abril de 2009

Resumo. *O objetivo deste trabalho de pesquisa em nível de mestrado é extrair informação em artigos científicos completos sobre efeitos relacionados ao domínio da doença Anemia Falciforme (AF). Para isso, primeiramente é necessário converter os artigos inerentemente no formato não-estruturado (PDF) para o formato semi-estruturado (XML), permitindo acesso aos níveis estruturais dos artigos por: página, parágrafo e sentença. Em seguida, a partir do acesso ao documento XML, será possível processar o texto a fim de identificar os efeitos (complicação e benefício) da AF originados a partir de um tratamento. Nesse contexto será proposta uma metodologia de pré-processamento textual, utilizando a combinação de três abordagens para extrair informação no domínio biomédico:* ***abordagem baseada em aprendizado de máquina****, utilizada para classificar as sentenças em complicação, benefício e outros (sentenças que não são complicação nem benefício), as quais são processadas do documento XML;* ***abordagem baseada em dicionário****, utilizada para identificar diretamente efeitos da AF nas sentenças classificadas; e* ***abordagem baseada em regras****, utilizada para identificar padrões de extração de efeitos com expressões regulares. Um desafio na extração de informação no contexto deste trabalho é lidar com um grande volume de dados. Assim, surge a oportunidade de utilizar a* ***Mineração de Textos*** *para processar arquivos em formato não-estruturado, identificando padrões textuais que serão armazenados no formato estruturado em um banco de dados relacional, para ser posteriormente utilizado por algoritmos de Mineração de Dados.*

Palavras-chave. *Banco de Dados Textual, Mineração de Textos, Pré-processamento, Extração de Informação, Domínio Biomédico e Doença Anemia Falciforme.*

[1] Os autores agradecem o apoio financeiro das seguintes agências de fomento à pesquisa do Brasil: CAPES, CNPq, FINEP e FAPESP.

1. Introdução

As informações relevantes estão mais em formato textual do que em imagens, gráficos, arquivos de música e vídeo ou até mesmo em equações. Segundo Tan [1] e Chen [2], 80% das informações, respectivamente, das empresas e de conteúdo on-line do mundo estão em documentos textuais. Estudos têm revelado que entre 80% e 98% de todos os dados disponíveis nos computadores consistem em documentos não-estruturados ou semi-estruturados, como e-mails, páginas HTML, arquivos PDF e muitos outros documentos textuais [3]. Além disso, a quantidade de informação disponível eletronicamente está aumentando consideravelmente nos últimos anos [4].

Diante da imensa quantidade de informação disponível em formato textual, os seres humanos não são capazes de processar (i.e., ler e assimilar) toda essa informação. Nesse contexto, a Mineração de Textos é uma tendência que favorece o processamento de textos, viabilizando o acesso a um grande volume de dados. Mineração de Textos [1], também conhecida como Descoberta de Conhecimento Textual [5] ou Mineração de Dados Textuais [6], refere-se ao processo de extrair informações úteis em documentos no formato textual não-estruturado através da identificação de conhecimento e exploração de padrões.

Este trabalho objetiva utilizar a Mineração de Textos para identificar e extrair informações novas, úteis e interessantes em artigos científicos do domínio biomédico, mais especificamente sobre a Anemia Falciforme que é uma doença genética e hereditária considerada como problema de saúde pública no Brasil [7].

As informações a serem extraídas estão em artigos completos ao invés de resumos. Os benefícios na extração de informação em artigos completos sobressaem à extração em resumos, não obstante se deparar com outros diferentes problemas, como conversão de formatos, proteções de *copyright* e maior tempo de processamento [8, 9].

A informação pontual a ser extraída é sobre efeitos (complicação e benefício) dessa doença, originados a partir de um tratamento. A finalidade da extração é organizar e armazenar essas informações em um Banco de Dados Relacional para viabilizar uma posterior Mineração de Dados, a fim de encontrar relacionamento de informação desconhecida.

Para alcançar esse objetivo é proposta uma metodologia de pré-processamento textual combinando três abordagens predominantes na literatura biomédica: aprendizado de máquina, regras e dicionário. Considera-se como hipótese deste trabalho a possibilidade de extrair informações da doença mencionada para auxiliar o médico e complementar seu conhecimento atual. As informações a serem extraídas encontram-se em artigos científicos no formato XML pré-processado originalmente do arquivo PDF.

Este artigo está estruturado como segue. Na Seção 2, são apresentadas as abordagens encontradas na literatura para extrair informação, enquanto, na Seção 3, são resumidos os trabalhos correlatos que utilizam essas abordagens. Na Seção 4, é proposta a metodologia de pré-processamento textual e, na Seção 5, são feitas algumas considerações finais.

2. Abordagens para Extração de Informação

Kevin Cohen e Hunter [10] apresentam duas abordagens para extração de informação: abordagem baseada em regras e baseada em aprendizado de máquina. A primeira faz o uso de algum tipo de conhecimento; a segunda utiliza classificadores para separar sentenças ou documentos. Krauthammer e Nenadic [11] e Ananiadou e McNaught [12] apresentam uma terceira abordagem, além dessas duas anteriores: abordagem baseada em dicionário que utiliza informações de um dicionário para auxiliar na identificação dos termos ou das entidades no texto. Essas abordagens são as três predominantes para extração de conhecimento no domínio biomédico.

A **abordagem baseada em dicionário** tem a vantagem de armazenar informações relacionadas a um determinado domínio e possibilitar a identificação de termos como nomes de gene e proteína. Alguns problemas dessa abordagem são: a limitação de nomes presentes

no dicionário, as variações de nome geram uma baixa revocação e os nomes curtos geram falsos positivos, o que diminui a precisão [13].

A **abordagem baseada em regras** tem algumas desvantagens: prolonga significativamente a construção de sistemas, reduz a capacidade de adaptação de regras em outro sistema e exclui termos que não correspondem aos padrões predefinidos. Tem, em geral, um desempenho melhor do que outras abordagens, no entanto há o problema de adaptação para novos domínios e classes [12].

As vantagens de se utilizar a **abordagem baseada em aprendizado de máquina** são a independência de domínio e a alta qualidade na predição. Os principais problemas relacionados aos algoritmos de aprendizado de máquina são a necessidade de grandes quantidades de dados de treinamento e a necessidade de novos treinamentos com o advento de novos dados. Em geral, a classificação é prejudicada quando o conjunto de dados de uma classe é pequeno (classe minoritária) em relação a outras classes [12].

Nota-se, portanto, que cada abordagem tem suas vantagens e desvantagens. Para desfrutar das características positivas dessas abordagens, surge a necessidade de se utilizar as suas melhores características.

3. Trabalhos Correlatos

Na literatura biomédica encontram-se trabalhos que extraem informação de resumos ou artigos completos predominantemente sobre gene ou proteína, utilizando a combinação das três abordagens discutidas anteriormente. A maioria extrai informação de resumos do MEDLINE e tem a característica de utilizar [14, 15, 16, 17] ou não utilizar [18, 19, 20, 21] um etiquetador de classes gramaticais, conhecido como *Part-Of-Speech tagger* (POS).

Os trabalhos que extraem informação em artigos completos têm objetivos diversos: extrair informação [22, 23], povoar um banco de dados [8, 24] ou destacar as sentenças de acordo com a consulta do usuário [25]. Ademais, alguns trabalhos utilizam POS [8, 22, 23], enquanto outros não utilizam um POS [24, 25].

Dentre os trabalhos, [23] obteve o maior percentual de precisão (72,5%) e revocação (50,7%), utilizando a combinação de aprendizado de máquina, dicionário e regras. Todavia, utilizando as mesmas técnicas na extração de informação em resumos obteve um percentual maior, respectivamente, de 85,7% e 66,7% [22]. Essa diferença demonstra as particularidades em extrair informação em artigos completos. Segundo Aaron Cohen e Hersh [26], o ABGene [22, 23] é uma das abordagens baseada em regras mais bem-sucedidas para reconhecimento de gene e proteína em textos biomédicos.

4. Metodologia de Pré-processamento Textual

O objetivo deste trabalho de pesquisa em nível de mestrado é atuar na segunda fase do processo de Mineração de Textos (i.e., na fase de pré-processamento, que vem após a fase de coleta de documentos) que, segundo Feldman e Sanger [27], é uma das fases mais críticas. Nesse sentido, a principal contribuição deste trabalho é a proposta de uma metodologia de pré-processamento textual, combinando três abordagens (i.e., aprendizado de máquina, regras e dicionário) para extrair informação no domínio biomédico (Figura 1).

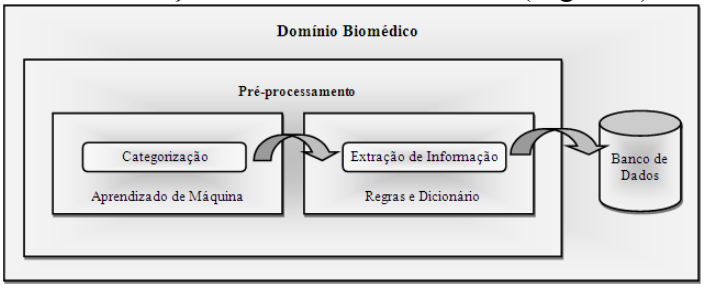

Figura 1. Pré-processamento utilizando categorização e extração de informação.

O aprendizado de máquina será utilizado para classificar as sentenças dos artigos científicos (tarefa de categorização). Em seguida, regras e dicionário serão desenvolvidos na tarefa de extração de informação: regras serão construídas a partir da análise de padrões identificados manualmente nas sentenças classificadas e serão utilizadas para aumentar a precisão da extração; o dicionário será construído manualmente a partir da identificação de novos termos encontrados nos artigos científicos e técnicas serão utilizadas para amenizar as variações de termos, a fim de aumentar a revocação da extração automática. Após o pré-processamento, as informações identificadas serão armazenadas em um banco de dados relacional.

4.1. Categorização

O Mover é uma ferramenta de classificação de aprendizado supervisionado de estruturas retóricas. Apresenta as organizações retóricas ou estruturas do texto, conhecidas como *moves,* usadas no texto, a fim de ajudar estudantes não-nativos que tenham dificuldades na leitura e escrita técnica, seja por falta de conhecimento ou experiência. O sistema aprende características estruturais de textos usando um número reduzido de exemplos de treinamento e pode ser aplicado em diferentes textos [28].

O classificador utilizado é o Naive Bayes, que utiliza a medida de utilidade *ganho de informação* para classificar os grupos de palavras de acordo com a pontuação. A representação de grupos (i.e., *bag of clusters*) foi desenvolvida para representar o conhecimento dos *moves* estruturais. O ruído é reduzido excluindo-se os grupos que estão abaixo de um certo limiar. O percentual de precisão alcançado por Anthony e Lashkia [28] nessa tarefa de auxiliar os estudantes foi de 86%. Esse resultado nos motivou a utilizar essa ferramenta a fim de classificar as sentenças relacionadas à doença Anemia Falciforme em três categorias (Figura 2): complicação, benefício e outros.

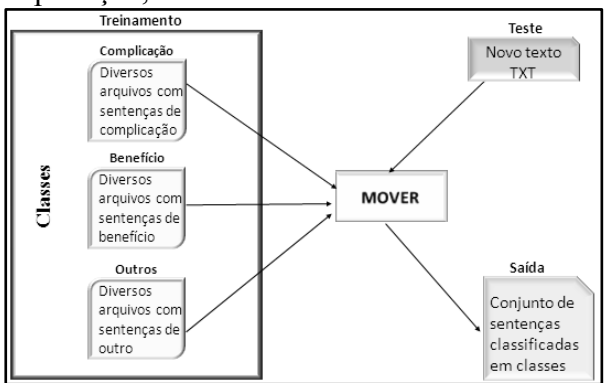

Figura 2. Processo de classificação: treinamento e teste na ferramenta Mover.

O processo de classificação é dividido em duas fases: treinamento e teste. Na primeira fase é necessário identificar as sentenças de alguns artigos (pretende-se selecionar entre 500 e 1.500 sentenças). Este processo é realizado de forma manual e contará com a ajuda de médicos especialistas da doença para validar se as sentenças foram classificadas corretamente. Após identificar as sentenças nos artigos, o classificador Mover será treinado com a quantidade de sentenças identificadas nos artigos. Em seguida, artigos com novas sentenças são selecionados para serem classificados pelo Mover.

4.2. Extração de Informação

Após a classificação das sentenças, ocorrerá a extração das informações relevantes dos artigos, mais especificamente das sentenças classificadas. Para isso será desenvolvido um módulo de extração de efeitos (complicação e benefício) da doença Anemia Falciforme, que faz parte da ferramenta SCAeXtractor (Sickle Cell Anemia eXtractor).

Este módulo utilizará as abordagens baseada em regras e dicionário. A primeira abordagem ajudará na identificação de alguns padrões com o auxílio da segunda. Para a

identificação dos padrões são utilizadas expressões regulares. Tem-se analisado algumas estratégias para:

1. Identificar verdadeiro positivo: algumas palavras correspondentes a complicação e benefício são armazenadas em um banco de dados auxiliar, a fim de confirmar as sentenças que foram realmente classificadas na devida classe;
2. Eliminar falso positivo: palavras relacionadas a fator de risco também são armazenadas para auxiliar na identificação de sentenças que foram classificadas erroneamente;
3. Elaborar outras estratégias para recuperar falsos negativos.

5. Considerações Finais

Este artigo apresentou a proposta de uma metodologia de pré-processamento textual para extrair informação de artigos científicos sobre a doença Anemia Falciforme. As contribuições teóricas são: a metodologia para extração de informação e o conhecimento das peculiaridades das informações do domínio biomédico. As contribuições práticas são: criação e disponibilização de recursos (coleção de documentos, dicionário e base de regras) e das ferramentas de extração de informação (SCAeXtractor) e de conversão e etiquetagem de formatos (SCAtRanslator - Sickle Cell Anemia tRanslator).

Atualmente, as ferramentas SCAeXtractor e SCAtRanslator estão em fase de implementação. Para avaliar o resultado gerado pelo classificador e extrator serão utilizadas, respectivamente, a taxa de precisão, e as medidas de precisão, revocação e Medida-F. A medida de concordância Kappa [29] será calculada a fim de conhecer o quão bem definido é a tarefa de classificação realizada pelos humanos. Essa medida será utilizada para comparar com o resultado obtido pelo classificador. Por fim, intenciona-se realizar quatro experimentos com o intuito de descobrir qual das combinações de técnicas de extração de informação possui a melhor qualidade, a saber: **1)** Somente regras geradas com expressão regular; **2)** Regras geradas com expressão regular com auxílio do dicionário; **3)** Aprendizado de máquina e regras geradas com expressão regular; **4)** Aprendizado de máquina e regras geradas com expressão regular com auxílio do dicionário.

Apesar do domínio deste trabalho se limitar à doença Anemia Falciforme, a metodologia proposta está sendo desenvolvida visando uma possível aplicação a outros domínios, sendo necessárias nesse caso algumas modificações nas regras e no dicionário para se adaptar ao novo domínio.

Referências

[1] TAN, A.-H. Text mining: the state of the art and the challenges. In: KDAD, Beijing, China. PAKDD, 1999. p. 71-76.

[2] CHEN, H. **Knowledge management systems:** a text mining perspective. Tucson, AZ: University of Arizona, 2001.

[3] CHEUNG, C. F.; LEE, W. B.; WANG, Y. A multi-facet taxonomy system with applications in unstructured knowledge management. **Journal of Knowledge Management**, v. 9, n. 6, p. 76-91, 2005.

[4] GANTZ, J. F. et al. **The expanding digital universe:** a forecast of worldwide information growth through 2010. IDC Whitepaper, 2007.

[5] FELDMAN, R.; DAGAN, I. Knowledge discovery in textual databases (KDT). In: KDD, Montréal, Québec. Menlo Park, CA: AAAI, 1995. p. 112-117.

[6] HEARST, M. A. Untangling text data mining. In: ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS, 37th, College Park, Maryland. Morristown, NJ: ACL, 1999. p. 3-10.

[7] BATISTA, A.; ANDRADE, T. C. Anemia falciforme: um problema de saúde pública no Brasil. **Universitas: Ciências da Saúde**, v. 3, n. 1, p. 83-99, 2005.

[8] CORNEY, D. P. A. et al. BioRAT: extracting biological information from full-length papers. **Bioinformatics**, v. 20, n. 17, p. 3206-3213, 2004.
[9] SCHUEMIE, M. J. et al. Distribution of information in biomedical abstracts and full-text publications. **Bioinformatics**, v. 20, n. 16, p. 2597-2604, 2004.
[10] COHEN, K. B.; HUNTER, L. Getting started in text mining. **PLoS Computational Biology**, v. 4, n. 1, p. 1-3, 2008.
[11] KRAUTHAMMER, M.; NENADIC, G. Term identification in the biomedical literature. **Journal of Biomedical Informatics**, v. 37, n. 6, p. 512-526, 2004.
[12] ANANIADOU, S.; MCNAUGHT, J. (Ed.). **Text mining for biology and biomedicine**. Norwood, MA: Artech House, 2006.
[13] TSURUOKA, Y.; TSUJII, J. I. Improving the performance of dictionary-based approaches in protein name recognition. **Journal of Biomedical Informatics**, v. 37, n. 6, p. 461-470, 2004.
[14] MIKA, S.; ROST, B. Protein names precisely peeled off free text. **Bioinformatics**, v. 20, p. i241-247, 2004. Suppl. 1.
[15] _____. NLProt: extracting protein names and sequences from papers. **Nucleic Acids Research**, v. 32, p. 634-637, 2004. Suppl. 2.
[16] ZHOU, G. et al. Recognizing names in biomedical texts: a machine learning approach. **Bioinformatics**, v. 20, n. 7, p. 1178-1190, 2004.
[17] CHUN, H.-W. et al. Extraction of gene-disease relations from medline using domain dictionaries and machine learning. In: PSB, 11th, Hawaii. 2006. p. 4-15.
[18] LEONARD, J. E.; COLOMBE, J. B.; LEVY, J. L. Finding relevant references to genes and proteins in Medline using a Bayesian approach. **Bioinformatics**, v. 18, n. 11, p. 1515-1522, Nov., 2002.
[19] SEKI, K.; MOSTAFA, J. An approach to protein name extraction using heuristics and a dictionary. In: ASIST, Long Beach, CA. 2003. p. 1-7.
[20] _____. A hybrid approach to protein name identification in biomedical texts. **Information Processing & Management**, v. 41, n. 4, p. 723-743, 2005.
[21] HANISCH, D. et al. ProMiner: rule-based protein and gene entity recognition. **BMC Bioinformatics**, v. 6, p. S14, 2005. Suppl. 1.
[22] TANABE, L.; WILBUR, W. J. Tagging gene and protein names in biomedical text. **Bioinformatics**, v. 18, n. 8, p. 1124-1132, 2002.
[23] _____. Tagging gene and protein names in full text articles. In: WORKSHOP ON NATURAL LANGUAGE PROCESSING IN THE BIOMEDICAL DOMAIN, Phildadelphia, Pennsylvania. Morristown, NJ: ACL, 2002. p. 9-13.
[24] BREMER, E. G. et al. Text mining of full text articles and creation of a knowledge base for analysis of microarray data. In: KELSI, Milan, Italy. 2004. p. 84-95.
[25] GARTEN, Y.; ALTMAN, R. Pharmspresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. **BMC Bioinformatics**, v. 10, p. S6, 2009. Suppl. 2.
[26] COHEN, A. M.; HERSH, W. R. A survey of current work in biomedical text mining. **Briefings in Bioinformatics**, v. 6, n. 1, p. 57-71, 2005.
[27] FELDMAN, R.; SANGER, J. **The text mining handbook:** advanced approaches in analyzing unstructured data. NY: Cambridge University, 2007.
[28] ANTHONY, L.; LASHKIA, G. V. Mover: a machine learning tool to assist in the reading and writing of technical papers. **IEEE Transactions on Professional Communication**, v. 46, n. 3, p. 185-193, 2003.
[29] FLEISS, J. L. Measuring nominal scale agreement among many raters. **Psychological Bulletin**, v. 76, n. 5, p. 378-382, 1971.