JHI JOURNAL OF HEALTH INFORMATICS

EDITORIAL

Mineração de Texto em Saúde

Evandro Eduardo Seron Ruiz
Professor Associado, Departamento de Computação e Matemática, Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto - DCM-FFCLRP Universidade de São Paulo - USP, Ribeirão Preto (SP), Brasil

O termo "mineração de texto" refere-se a análise de dados e informações contidas em textos escritos em modo livre, ou como comumente chamamos, em textos escritos em linguagem natural, ou seja, sem campos estruturados.

Especificamente na área da saúde humana, uma das fontes mais conhecidas e utilizadas desta forma de dados, os textos livres, são os Registros Eletrônicos em Saúde, ou RES. Na última década, os avanços das tecnologias e métodos relacionados àa informação em saúde proporcionaram uma produção intensa de grandes volumes de dados pelos usuários de sistemas computacionais, dados esses que hoje são armazenados em RES. Além de prontuários digitais, de repositórios, os RES são hoje a principal ferramenta do moderno profssional de saúde que o auxilia desde o registro da história clínica do paciente até a prescrição de medicamentos, da solicitação de exames e encaminhamentos, entre outras tarefas. As retribuições ao sistema de saúde pelo uso desta tecnologia vêm não só pelo acesso rápido aos registros dos pacientes, mas também vêm da possibilidade de análise destes registros pelo computador. Quando esta análise provém de manuscritos escritos em texto livre chamamos de *mineração de texto*.

A mineração de texto (TM, do inglês *text mining*) é relevante hoje não apenas para a transformação em informação e conhecimentos dos dados textuais contidos nos RES mas também para expandir o conhecimento na grande área de biomedicina. A saúde humana é produzida por processos biológicos complexos, em dimensões físicas que vão da molécula aos tecidos e órgãos, além de se envolverem em escalas de tempo das mais diversas. Assim, para conhecermos melhor este complexo sistema, as aplicações de TM contemplam várias áreas de atuação, tais como: a anotação das funções gênicas e protéicas, a predição de interações entre protéinas e drogas, a caracterização de doenças através de associações gênicas e também através de associações a fenótipos e, inclusive, pela área de respostas automatizadas a perguntas biomédicas, área também conhecida pelo termo *question and answering*[1].

Uma das referências ao desenvolvimento de novas abordagens na área de mineração de textos em saúde, ou mais amplamente na área de Processamento de Língua Natural, são os desafios da i2B2, os *i2B2 Challenges* que ocorrem desde 2006. O *Informatics for Integrating Biology and the Bedside* (i2b2) é um centro de pesquisa financiado pelo NIH para fomentar o desenvolvimento dos EUA na infra-estrutura computacional para informática biomédica. Como exemplo de aplicação de TM, no *The i2b2/UT Health 2014 Cardiac Risk Factors Challenge* foram comparados os resultados de equipes que conseguiram extrair dos RES anotações sobre os fatores de risco cardíaco tais como a diabetes, a obesidade, a pressão alta e associação ou não ao tabagismo. Esperava-se também que estas informações fossem extraídas com o fator temporal de presença do risco desde a criação do registro. Neste desafio a equipe de Cormack[2] mostra como as metodologias de TM podem obter precisão na recuperação deste tipo de informação na casa de 90%.

Por alguns anos a mineração de textos apoiou-se fundamentalmente nos métodos estatísticos para obter novas informações e conhecimento. Os grandes volumes atuais de textos na área de saúde facilitam a aplicação de métodos que basicamente representam o texto, ou parte dele, como um vetor multidimensional, ou mesmo como uma matriz numérica[3]. Dada a natureza heterogênea da escrita, a extração automatizada da informação relevante em saúde não é uma tarefa trivial. Atualmente os pesquisadores aliam outras informações ao conhecimento estatístico, tais como as informações advindas da análise sintática de frases do texto, aliam também conhecimentos semânticos proporcionado por redes semânticas elaboradas manualmente, conhecimentos estabelecidos em ontologias, advindos de léxicos, de dicionários, de tesauros, entre outros.

Atualmente as frentes de oportunidades para o progresso desta área de pesquisa passam não só por formulações adequadas de métodos que agregam todos os tipos de informações que possam ser extraídas de um texto mas também exigem da comunidade acadêmica de 'padrões ouro' que possam ser usados para comparar o desempenho destes métodos. Na fronteira do conhecimento ainda estão os métodos de TM baseados em grafos e/ou redes complexas além dos auxiliados por abordagens topológicos ou pela aplicação de algoritmos evolutivos. Para países

que ainda não têm estabelecidas terminologias médicas padronizadas em suas línguas, códigos de doenças, procedimentos e classificações que atendam a sua realidade, cabe a eles estabelecerem centros de pesquisa sobre terminologias médicas que fomentem a criação de uma infra-estrutura lógica para o processamento da informação em saúde escrita na sua língua corrente.

REFERÊNCIAS

1. Andrade-Navarro M, Perez-Iratxeta C. Text mining of biomedical literature: Doing well, but we could be doing better. Methods. 2015;74:1 - 2. Text mining of biomedical literature. Available from: http://www.sciencedirect.com/science/article/pii/S1046202315000262.
2. Cormack J, Nath C, Milward D, Raja K, Jonnalagadda SR. Agile text mining for the 2014 i2b2/UTHealth Cardiac risk factors challenge. Journal of Biomedical Informatics. 2015;58, Supplement:S120 - S127. Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415001410.
3. Holzinger A, Schantl J, Schroettner M, Seifert C, Verspoor K. Biomedical text mining: State-of-the-art, open problems and future challenges. In: Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. Springer; 2014. p. 271-300.