Protótipo de Sistema Computacional para o Mapeamento de Formulários Médicos

Willian Zalewski[1,5], Huei Diana Lee[1,5], André Gustavo Maletzke[1,3,5], Newton Spolaôr[1,5], Ana Carolina Lorena[4], Adewole Marcus Jacob Freitas Caetano[4], João José Fagundes[2], Cláudio Saddy Rodrigues Coy[2], Feng Chung Wu[1,2,5]

[1]Laboratório de Bioinformática — LABI,
Universidade Estadual do Oeste do Paraná — UNIOESTE, Brasil
[2]Serviço de Coloproctologia da Faculdade de Ciências Médicas — FCM,
Universidade Estadual de Campinas — UNICAMP, Brasil
[3] Laboratório de Inteligência Computacional — LABIC,
Universidade de São Paulo — USP, Brasil
[4]Centro de Matemática, Computação e Cognição — CMCC,
Universidade Federal do ABC — UFABC, Brasil
[5]Parque Tecnológico Itaipu — PTI, Brasil

Introdução

Pesquisas relacionadas a determinadas enfermidades como a doença de Crohn [1], envolvem grandes volumes de dados. Nesse sentido, cada vez mais são aplicados processos computacionais que possam auxiliar na análise desses dados. Um dos processos computacionais que provê esse apoio é o de Descoberta de Conhecimento em Bases de Dados – DCBD [2]. Entretanto, na medicina os dados encontram-se frequentemente em formatos desestruturados ou semi-estruturados, por exemplo, em laudos médicos ou em formulários impressos, o que dificulta a análise direta por meio de métodos computacionais.

Desse modo, o objetivo deste trabalho é apresentar e avaliar uma metodologia e um protótipo de sistema computacional desenvolvidos para realizar o mapeamento automático de Formulários Médicos – FM, compostos por Campos de Múltipla Escolha – CME – e Campos Numéricos – CN –, para Bases de Dados – BD – estruturadas.

Materiais e Métodos

Em [3] é proposta uma metodologia, estruturada em três etapas, para auxiliar na coleta e armazenamento de dados, de modo estruturado, por meio de formulários médicos impressos. Na primeira etapa é construído um modelo de FM contendo CME e CN. É realizada também a construção da BD, para a qual as informações contidas nos FM serão mapeadas e é gerado o Arquivo de Interpretação – AI –, que indica o modo como os campos serão mapeados.

Na segunda etapa, o modelo de FM gerado é impresso e, posteriormente, digitalizado. Após, a imagem desse FM é submetida a algoritmos de Inteligência Artificial – IA –, os quais auxiliam na construção de uma Base de Padrões – BP – que compreende a localização dos campos no formulário.

Na terceira etapa são impressas cópias do modelo de FM, as quais são preenchidas pelos usuários. Posteriormente, essas cópias são digitalizadas e mapeadas para uma BD com o auxílio dos padrões definidos na BP e das regras que compõem o AI.

A partir dessa metodologia foi desenvolvido um protótipo de sistema computacional para integrar todas as etapas dessa metodologia. Esse sistema foi construído utilizando a linguagem JAVA[1].

As três etapas da metodologia foram projetadas e implementadas no protótipo em quatro módulos: (1) Módulo de Geração de Formulários, (2) Módulo de Construção de Padrões sobre os Formulários; (3) Módulo de Mapeamento de Formulários; (4) Módulo de Reconhecimento de Caracteres Manuscritos.

O módulo (1), Figura 1 (a) tem como objetivo implementar as tarefas correspondentes à primeira etapa da metodologia. Portanto, esse módulo é responsável pela construção do modelo de FM, da BD e do AI. Os formulários são gerados no formato *Portable Document Format* – PDF.

O módulo (2) engloba as tarefas realizadas na segunda etapa da metodologia. Na Figura 1(b) é apresentada a interface gráfica do módulo (2).

Para realizar o mapeamento dos CME em formulários preenchidos é utilizado o módulo (3), Figura 1 (c). Esse módulo permite identificar qual resposta de cada pergunta foi preenchida no formulário e mapear essas informações para a BD, por meio das informações que foram registradas na BP e no AI.

Ainda na interface gráfica da Figura 1 (c) é disponibilizado o módulo (4), que é utilizado para o reco-

[1]`http://www.java.sun.com`

nhecimento e mapeamento dos dígitos escritos nos CN. No desenvolvimento desse módulo foram estudadas, aplicadas e avaliadas as principais técnicas adotadas na literatura para o reconhecimento de caracteres manuscritos [4].

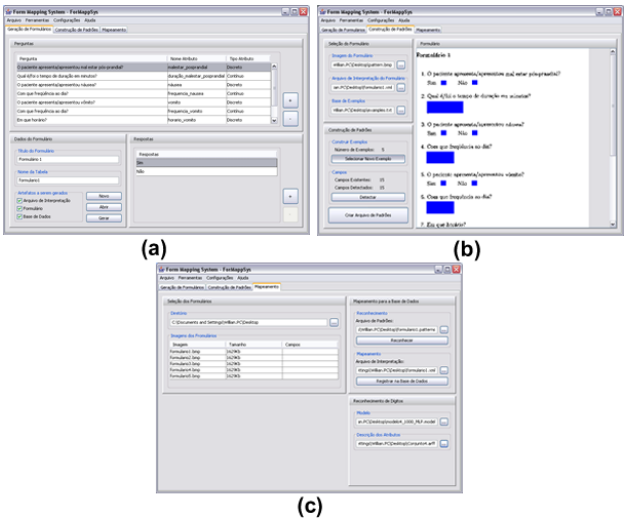

Figura 1: Interfaces Gráficas dos Módulos do Protótipo de Sistema Implementado.

Para a avaliação foi construído um modelo de FM, a partir de um protocolo de informações referentes à doença de Crohn, composto por cinco perguntas com CME e quatro perguntas com CN. Esse protocolo foi definido por especialistas e pesquisadores das instituições parceiras.

Esse FM foi impresso em escala de cinza utilizando uma impressora HP Color LaserJet 2605dn com 1200 *pixels* por polegada. Em seguida esse FM foi digitalizado por meio da utilização de um *scanner* HP ScanJet 5550c com configuração de 75 *pixels* por polegada, 50% de brilho e 50% de contraste.

O modelo de FM foi submetido ao módulo (1) para a construção da BP. Após, foram geradas e preenchidas 80 cópias, por oito colaboradores, de modo que cada um preencheu dez FM. Posteriormente, os 80 FM foram digitalizados seguindo a mesma configuração utilizada para a digitalização do modelo de FM.

Os CME foram preenchidos de modo *ad libitum* e para os CN foi recomendado o preenchimento dentro do espaço delimitado pela grade em tons de cinza e a escrita separada dos caracteres numéricos.

Resultados

A partir da avaliação experimental, observou-se que a identificação de informações em campos de múltipla escolha atingiu 94,25% de precisão. O reconhecimento de campos numéricos, considerando apenas os dígitos, atingiu uma precisão de 89,11% e o reconhecimento dos 320 campos numéricos como um todo, 253 campos foram mapeados adequadamente, o que correspondeu a uma precisão de 79,06%.

Discussão e Conclusões

A precisão da identificação de campos de múltipla escolha e do reconhecimento de campos numéricos foi considerada satisfatória por especialistas, embora seja necessário uma maior confiabilidade para aplicações em situações reais. Portanto, necessita-se atingir uma precisão maior para a sua aplicabilidade em ambientes médicos. A adoção de restrições durante a grafia dos dígitos poderia contribuir para a redução da variabilidade presente nos caracteres e, conseqüentemente, uma maior confiabilidade na classificação desses caracteres.

Trabalhos futuros incluem a aplicação de outros algoritmos para a classificação dos CN e a aplicação de técnicas de seleção de atributos relacionados às características dos dígitos grafados nos FM. Esses estudos poderão auxiliar na consolidação do protótipo de sistema, contribuindo para a aplicação de processos, como o de DCBD, para a extração de padrões que possam estar contidos na BD construída.

Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado (PDTA/FPTI – BR) pelo auxílio por meio da linha de financiamento de bolsas.

Referências

[1] Cordeiro F. Endoscopia Digestiva. Editora Média e Científica Ltda.; 1994.

[2] Fayyad UM, Platestsky-Shapiro G, Smyth P. From data mining to knowledge discovery: an overview. In: American Association for Artificial Intelligence; 1996. p. 1–30.

[3] Maletzke AG, Lee HD, Zalewski W, Edson T Matsubara RFV, Coy CSR, Fagundes JJ, et al. Mapeamento de Informações Médicas descritas em Formulários para Bases de Dados Estruturadas. In: VII Workshop de Informática Médica. Porto de Galinhas, PE, Brasil; 2007. p. 49–58.

[4] Arica N, Yarman-Vural FT. An overview of character recognition focused on off-line handwriting. IEEE Transactions on Systems Man and Cybernetics Part C: Applications and Reviews. 2001;31:216–233.

Contato

W. Zalewski — willzal@gmail.com
Laboratório de Bioinformática — LABI, Universidade Estadual do Oeste do Paraná — UNIOESTE, Parque Tecnológico Itaipu — PTI, Av. Tancredo Neves, 6731, CEP 85866-900, Foz do Iguaçu — PR.