Um Sistema para o Mapeamento de Informações Contidas em Formulários Médicos para Bases de Dados Estruturadas

Willian Zalewski[1], Huei Diana Lee[1], André Gustavo Maletzke[2]
Cláudio Sady Rodrigues Coy[3], João José Fagundes[3], Feng Chung Wu[1,3]

[1]Centro de Engenharias e Ciências Exatas – Universidade Estadual do Oeste do Paraná
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Caixa Postal 39, 85856-970 – Foz do Iguaçu, PR, Brasil

[2]Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo
Laboratório de Inteligência Computacional – LABIC
Caixa Postal 668, 13560-970 – São Carlos, SP, Brasil

[3]Faculdade de Ciências Médicas – Universidade Estadual de Campinas
Serviço de Coloproctologia
Caixa Postal 6111, 13083-970 – Campinas, SP, Brasil

{willzal,hueidianalee,andregustavom}@gmail.com

Abstract. *Focused in helping researches related to Crohn disease, it was developed a protocol of information considered relevant by experts. Because of the great quantity of attributes defined on this protocol, a methodology and a computational system were developed to automatically map data contained in printed medical forms filled manually. These forms were structured in a questions and answers layout, composed by multiple choice and numeric fields. An experimental evaluation of this methodology using fifty forms was held in this work. The mapping performed by the system presented 99.6% precision in multiple choice fields and 90.25% in numeric fields. This way, the needed time and subjectivity were eliminated or minimized on the mapping process.*

Resumo. *Com o objetivo de auxiliar em pesquisas relacionadas à doença de Crohn e devido à grande quantidade de atributos definidos nesse protocolo foi desenvolvida uma metodologia e um sistema computacional para realizar o mapeamento automático de dados contidos em formulários médicos impressos preenchidos manualmente. Esses formulários foram estruturados no formato de perguntas e respostas. Neste trabalho foi realizada uma avaliação experimental dessa metodologia utilizando 50 formulários. O mapeamento realizado pelo sistema apresentou uma precisão de 99,60% para os campos de múltipla escolha e 90.25% para os campos numéricos. Desse modo, o custo de tempo e a subjetividade foram eliminados ou minimizados durante o mapeamento.*

1. Introdução

A análise de dados é uma tarefa que cada vez mais tem sido aplicada em diversas áreas com o intuito de auxiliar no processo de tomada de decisão. Todavia, o aumento do volume de dados armazenados tem tornado essa tarefa crescentemente complexa por meio

de abordagens tradicionais e/ou manuais. Nesse sentido, cada vez mais métodos e processos computacionais têm sido propostos e aplicados na análise de grandes conjuntos de dados. Dentre esses processos, o de Descoberta de Conhecimento em Bases de Dados — DCBD — [1] tem sido utilizado como apoio na tarefa de análise de dados. Esse processo tem como objetivo a extração de padrões contidos nos dados, de modo que esses padrões constituam uma fonte de informação interessante e relevante para especialistas de diversos domínios.

Para que o processo de DCBD possa ser aplicado é necessário que os dados estejam dispostos em um formato estruturado como a representação atributo-valor [2, 3]. No entanto, geralmente os dados encontram-se em formatos desestruturados ou semi-estruturados. Na área de medicina, freqüentemente, os dados estão dispostos em laudos médicos e formulários impressos contendo informações como histórico do paciente e sintomatologia, fato este que dificulta a análise direta por meio de métodos computacionais. Diversos motivos estão relacionados à indisponibilidade desses dados em um formato adequado, dentre os quais a não existência de computadores em ambulatórios médicos, a consideração por muitos profissionais da área de medicina de que a utilização de documentos impressos torna o relacionamento com o paciente menos impessoal e a necessidade de se manter registros impressos. Portanto, para que processos, como o DCBD, possam ser aplicados é necessário representar esses dados em formato estruturado, como a representação atributo-valor.

Um dos temas que tem despertado grande interesse entre os pesquisadores da área médica está relacionado às doenças inflamatórias intestinais, como a doença de Crohn. Nessa doença, células imunologicamente ativas agridem o aparelho digestivo, da boca até o ânus, em especial a parte inferior do intestino delgado (ílio) e do intestino grosso (cólon), provocando graves lesões como esfoliação, diarréia, aumento da velocidade do trânsito intestinal, dificuldade para absorver os nutrientes e enfraquecimento [4].

Atualmente, ainda não se conhece a causa exata da doença de Crohn. Inúmeras pesquisas tentaram relacionar fatores genéticos, ambientais, alimentares e infecções como responsáveis pela doença. Porém, nenhum desses fatores, isoladamente, conseguiu explicar a manifestação e o desenvolvimento dessa doença [5]. Nesse contexto, processos computacionais como o de DCBD, poderiam ser úteis como ferramentas de apoio para pesquisas relacionadas a essa doença. Isso permitiria que dados de acompanhamento de pacientes juntamente com os resultados de tratamentos pudessem ser utilizados para a construção de modelos e a extração de padrões sobre a doença.

O objetivo deste trabalho consiste em apresentar e avaliar a metodologia e o sistema computacional desenvolvidos para realizar o mapeamento automático de informações contidas em formulários médicos.

O restante deste trabalho está organizado do seguinte modo: na Seção 2 é apresentada a metodologia proposta para o mapeamento automático de dados contidos em formulários médicos impressos para bases de dados; na Seção 3, o sistema computacional ForMappSys é descrito por meio dos módulos e das funcionalidades que o compõem; a avaliação experimental da metodologia e do sistema computacional é apresentada na Seção 4 e os resultados na Seção 5; as conclusões e trabalhos futuros são apresentados na Seção 6.

2. Metodologia para o Mapeamento Automático de Formulários

Este trabalho está inserido no projeto de Análise Inteligente de Dados aplicado ao Mapeamento de Dados — AidMD — [6, 7, 8], o qual está sendo desenvolvido pelo Laboratório de Bioinformática — LABI — da Universidade Estadual do Oeste do Paraná — UNIOESTE — em conjunto com o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade Estadual de Campinas — UNICAMP — e o Laboratório de Inteligência Computacional — LABIC — da Universidade de São Paulo — USP —, São Carlos. Dentro desse projeto estão sendo desenvolvidos outros projetos, dentre os quais o de mapeamento de formulários médicos impressos de múltipla escolha para Bases de Dados — BD — estruturadas, ao qual este trabalho esta relacionado. Em [8] é apresentada uma metodologia aplicada ao mapeamento desses formulários, a qual está organizada em três etapas (Figura 1): (1) Geração de Formulários e Construção da BD, (2) Construção de Padrões sobre Formulários e (3) Mapeamento de Formulários e Preenchimento da BD.

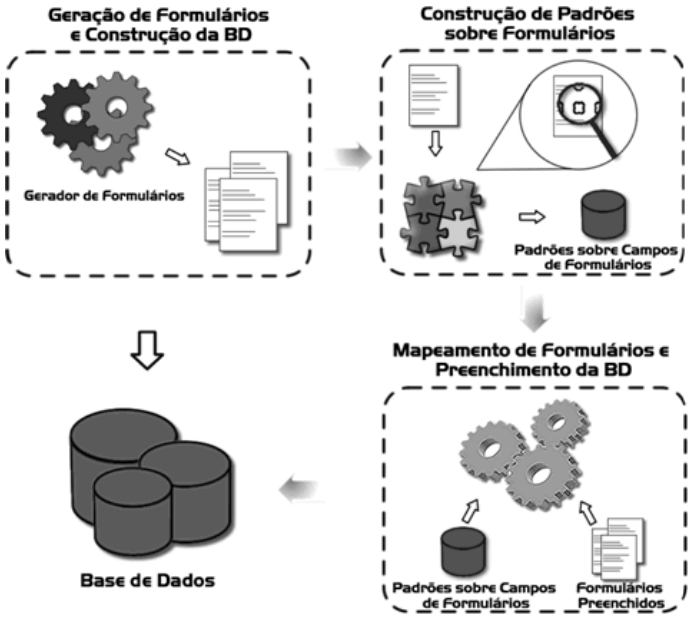

Figura 1. Metodologia Desenvolvida para o Mapeamento de Formulários Impressos [8]

Na primeira etapa, devido à grande quantidade de atributos que deverão ser mapeados para a BD, os quais foram definidos por meio de reuniões com especialistas, torna-se necessário elaborar um método eficiente e menos dispendioso para realizar essa tarefa. Portanto, nessa etapa são construídos formulários de múltipla escolha a partir dos atributos definidos anteriormente, possibilitando realizar marcações nos campos desses formulários para selecionar o valor observado para cada atributo. Ainda outras características estão presentes no formulário como marcas de referência e identificação do formulário, necessárias nas próximas etapas. Na segunda etapa, o formulário gerado na etapa anterior é digitalizado e submetido a algoritmos de Inteligência Artificial — IA —, os quais auxiliam na construção de padrões sobre esse formulário. É importante ressaltar que essa etapa torna o processo de mapeamento de diferentes formulários, seja por ruídos causados no processo de digitalização ou pela distribuição dos campos, mais robusto e eficiente. Na terceira etapa cópias preenchidas do formulário são mapeadas para a BD por meio dos padrões gerados na etapa anterior.

3. Sistema Computacional — *ForMappSys*

A partir da metodologia e do protótipo desenvolvidos em [8, 9] foi construído o *Form Mapping System* — ForMappSys —, o qual consiste em um aplicativo computacional desenvolvido com o objetivo de integrar todas as etapas da metodologia de mapeamento de formulários. O sistema ForMappSys foi desenvolvido em linguagem JAVA 5 SE[1] utilizando o ambiente integrado de desenvolvimento NetBeans IDE 5.5.1[2]. O desenvolvimento foi realizado baseando-se no paradigma de programação orientado a objetos. As três etapas da metodologia para o mapeamento de formulários foram elaboradas e implementadas em quatro módulos dentro do sistema ForMappSys, os quais são:

- Módulo de Geração de Formulários;
- Módulo de Construção de Padrões sobre os Formulários;
- Módulo de Mapeamento de Formulários;
- Módulo de Reconhecimento de Caracteres Manuscritos.

3.1. Módulo de Geração de Formulários

Esse módulo tem por objetivo implementar as tarefas correspondentes à etapa de Geração de Formulários e Construção da BD. Nessa etapa são construídos os formulários a partir de um conjunto de perguntas e respostas, uma BD para a qual essas respostas são mapeadas e um Arquivo de Interpretação — AI —, o qual contém as regras de preenchimento da BD para cada tipo de formulário construído. O AI está representado na linguagem *Extensible Markup Language* — XML[3] — e é responsável por indicar o valor que deve ser armazenado na BD de acordo com as respostas que foram marcadas no formulário [8].

Os formulários construídos estão estruturados no formato de perguntas e respostas, sendo que uma determinada pergunta pode conter várias respostas, caso seja de múltipla escolha, ou um único campo caso seja uma pergunta de resposta numérica. O formato desses campos no formulário e modo de preenchimento dos mesmos é apresentado na Figura 2. Cada pergunta representa um atributo na BD e a resposta preenchida, o valor que é conferido a este atributo na BD.

Após a definição das perguntas e suas respectivas respostas, a construção dos formulários é realizada de maneira automática por meio da geração de um arquivo na linguagem LaTeX[4], o qual é processado pelo aplicativo PDFLaTeX, presente no conjunto de ferramentas disponibilizadas pelo MikTeX[5], gerando o formulário no formato *Portable Document Format* — PDF [10].

Os formulários gerados possuem uma Marca de Referência — MR —, representada por uma linha horizontal, localizada no cabeçalho do formulário e que se estende por quase a totalidade da largura do formulário. Essa MR é utilizada em etapas posteriores da metodologia para auxiliar na correção de imperfeições dos formulários e na localização dos campos de preenchimento.

Esse módulo disponibiliza uma interface gráfica com o usuário que possibilita a edição de perguntas, respostas e informações que irão compor o formulário. Também é

[1] http://www.java.sun.com
[2] http://www.netbeans.org
[3] http://www.w3c.org/XML
[4] http://www.latex-project.org
[5] http://www.miktex.org

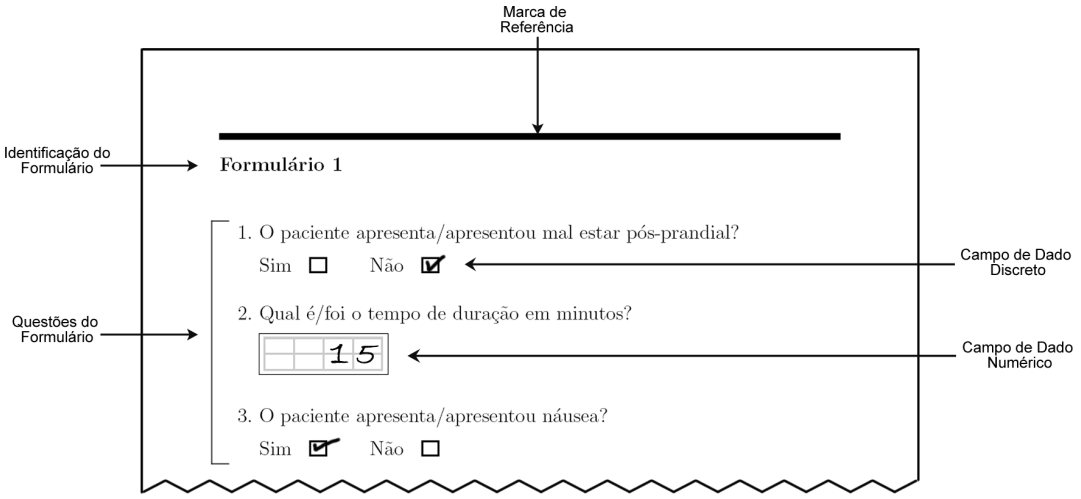

Figura 2. Representação Parcial de um Formulário Gerado pelo Sistema For-MappSys

possível indicar configurações sobre esses dados em relação à BD na qual serão armazenados e selecionar quais artefatos serão gerados. Essa interface gráfica é apresentada na Figura 3.

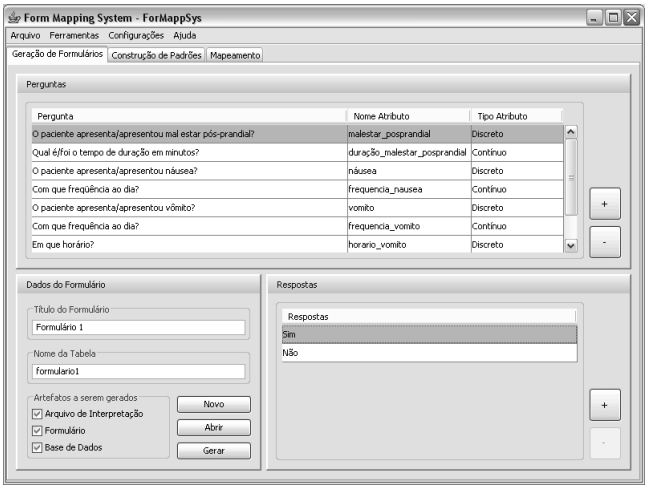

Figura 3. Interface Gráfica do Módulo de Geração de Formulários

Como mencionado, nessa etapa também é realizada a construção da BD, a qual foi implementada utilizando o Sistema de Gerenciamento de Banco de Dados MySQL[6].

3.2. Módulo de Construção de Padrões sobre os Formulários

Nesse módulo foram implementadas as funcionalidades referentes à etapa de Construção de Padrões sobre Formulários. Nessa etapa o objetivo é mapear a localização dos campos a partir de um formulário modelo (não preenchido) e armazenar essas informações para que possam ser utilizadas na etapa de Mapeamento de Formulários e Preenchimento da BD. O registro das informações sobre a localização dos campos do formulário é realizado em um arquivo denominado Base de Padrões [9].

[6]http://www.mysql.com

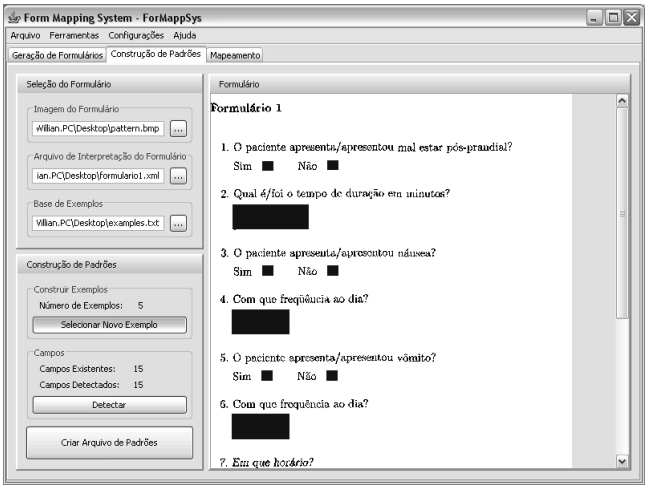

Figura 4. Interface Gráfica do Módulo de Construção de Padrões sobre os Formulários

Na Figura 4 é apresentada uma ilustração da tela do sistema ForMappSys referente ao Módulo de Construção de Padrões sobre Formulários. Por meio dessa interface gráfica é possível visualizar a imagem do formulário modelo, as informações como a quantidade de campos de exemplos escolhidos, a quantidade de campos existentes no formulário, o número de campos que já foram reconhecidos no formulário e selecionar os campos que deverão ser mapeados. Desse modo, quando todos os campos são identificados, o sistema habilita a opção para a criação da Base de Padrões.

3.3. Módulo de Mapeamento de Formulários

As operações implementadas nesse módulo correspondem aos requisitos da etapa de Mapeamento de Formulários e Preenchimento da BD. O objetivo dessa etapa é identificar qual resposta de cada pergunta foi preenchida no formulário e mapear essas informações para a BD. Para isso são utilizadas as informações que foram registradas na Base de Padrões, a qual foi construída na etapa anterior da metodologia. O Arquivo de Interpretação, gerado na primeira etapa da metodologia, é utilizado nesse momento para indicar de acordo com o campo preenchido qual será o valor a ser registrado na BD. O tratamento de campos numéricos é realizado pelo Módulo de Reconhecimento de Caracteres Manuscritos.

Para o Módulo de Mapeamento de Formulários foi elaborada uma interface gráfica (Figura 5), a qual permite selecionar as imagens dos formulários que deverão ser mapeados para a BD, a Base de Padrões e o Arquivo de Interpretação que serão utilizados. Também são disponibilizadas funcionalidades para a seleção do tipo de modelo de classificação que será utilizado pelo Módulo de Reconhecimento de Caracteres Manuscritos.

3.4. Módulo de Reconhecimento de Caracteres Manuscritos

O mapeamento de dados numéricos contidos nos formulários foi possível por meio do desenvolvimento do Módulo de Reconhecimento de Caracteres Manuscritos. Para a construção desse módulo, foram pesquisadas as principais técnicas adotadas na literatura para a extração de características e construção de modelos para o reconhecimento de

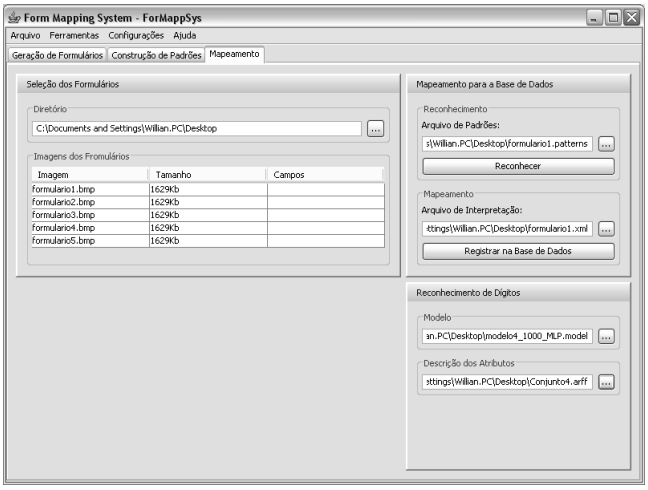

Figura 5. Interface Gráfica do Módulo de Mapeamento de Formulários

caracteres manuscritos [11, 12, 13]. A partir desse estudo foram avaliadas diversas dessas técnicas de extração de características de modo individual e combinadas, bem como distintos modelos [14]. Com base nos resultados dessas avaliações, neste trabalho foi construído um modelo de classificação, o qual consiste em uma Rede Neuronal Artificial — RNA [15]. As RNAs consistem em um método de solucionar problemas de IA por meio da construção de um modelo que simula o comportamento do cérebro humano. São técnicas computacionais que apresentam um modelo inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento por meio da experiência [16]. Neste trabalho foi utilizada a ferramenta WEKA[7], considerando os parâmetros com valores padrão, para a construção do modelo de classificação.

4. Avaliação Experimental

Por meio da utilização do sistema ForMappSys foi construído um formulário modelo composto por cinco perguntas com respostas de múltipla escolha e quatro perguntas com respostas numéricas. Dentre as perguntas com respostas de múltipla escolha, quatro são definidas segundo dois atributos (Sim e Não) e uma possui três possíveis respostas (Antes das refeições, Durante e Após as refeições). Em relação às perguntas com respostas numéricas foram gerados três campos com capacidade de escrita de três caracteres e um campo com capacidade para quatro caracteres. O conteúdo utilizado na elaboração dos formulários foi definido durante reuniões realizadas com especialistas da área de medicina.

O formulário modelo, não preenchido, foi utilizado para a construção de padrões referentes à localização dos campos.

Esse formulário foi impresso em escala de cinza utilizando uma impressora HP Color LaserJet 2605dn com 1200 pixels por polegada. Posteriormente, o formulário modelo 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 formulário modelo em formato digital, não preenchido, foi submetido ao sistema ForMappSys para

[7]http://www.cs.waikato.ac.nz/ml/weka

a construção da Base de Padrões (localização dos campos). Posteriormente, a partir do formulário modelo foram geradas 50 cópias, as quais foram preenchidas por dez colaboradores, de modo que cada um preencheu cinco formulários. Após a realização do preenchimento, os 50 formulários foram digitalizados seguindo a mesma configuração utilizada para a digitalização do formulário modelo.

O preenchimento de todos os formulários foi realizado utilizando caneta esferográfica de cor preta. Para os campos de múltipla escolha o preenchimento foi realizado de modo bastante livre, apenas fixando-se como critério a realização do preenchimento próximo ao centro de cada marca, aplicando-se uma pressão normal de escrita. Em relação aos campos numéricos, foi recomendado o preenchimento dos caracteres dentro do espaço delimitado pela grade em tons de cinza, contida no campo numérico, e que os caracteres fossem escritos separadamente, de modo que não existisse conexão entre os mesmos. Foi solicitado também que a escrita de alguns caracteres obedecessem a determinadas propriedades como a existência de uma base horizontal na região inferior do caractere 1, de um orifício na parte inferior do caractere 2 e de uma linha horizontal na porção média do caractere 7. Essas propriedades podem ser visualizadas na Figura 6, na qual são apresentados os caracteres 1, 2 e 7 com as características mencionadas.

Figura 6. Propriedades dos caracteres 1, 2 e 7 utilizados no experimento

5. Resultados e Discussão

Após a utilização do sistema ForMappSys para realizar o mapeamento dos 50 formulários, a Base de Dados preenchida foi analisada para avaliar o desempenho da metodologia utilizando o sistema. Considerando todos os formulários, em relação às 250 perguntas com respostas de múltipla escolha, somente uma não foi mapeada corretamente. Desse modo, pode-se constatar que o mapeamento dos campos de múltipla escolha obteve 99,60% de precisão. Em relação ao reconhecimento dos 503 caracteres preenchidos nos campos numéricos, entre todos os formulários, foi constatada uma precisão de 90.25%.

Com base na análise dos resultados obtidos nessa avaliação experimental, observou-se que grande parte dos caracteres preenchidos nos formulários foram classificados corretamente, no entanto, para a aplicação do sistema em situações reais é necessária uma precisão global ainda melhor.

Um fator determinante no sucesso de aplicações que utilizam técnicas de reconhecimento de caracteres manuscritos refere-se ao estabelecimento de restrições, como a definição de recomendações mais restritas quanto à escrita dos números e marcação dos campos. Desse modo, é possível reduzir a grande variabilidade de parâmetros existentes nesse tipo de aplicação, permitindo uma maior confiabilidade em relação à veracidade dos dados.

6. Conclusão e Trabalhos Futuros

Os resultados obtidos neste trabalho para a avaliação do sistema ForMappSys foram considerados satisfatórios pelos especialistas do domínio, de modo que o custo de tempo e

a subjetividade foram eliminados ou minimizados durante o mapeamento. No entanto, apesar do bom desempenho em relação ao mapeamento dos campos numéricos, para a aplicação do sistema em situações reais, é necessário um maior nível de confiabilidade.

Trabalhos futuros incluem a aplicação da metodologia considerando amostras de dados mais amplas e o desenvolvimento de outras técnicas para o reconhecimento de caracteres manuscritos, bem como a avaliação de outros classificadores com o intuito de desenvolver modelos ainda mais confiáveis para a tarefa de classificação de dígitos manuscritos. Outro trabalho futuro inclui a aplicação de métodos computacionais para a extração de padrões que possam estar contidos na Base de Dados, relacionada à doença de Crohn, construída por meio da metodologia de mapeamento de formulários implementada pelo sistema ForMappSys. Essa metodologia poderá ser aplicada também a outros domínios.

Agradecimentos

Ao Programa de Desenvolvimento Tecnológico Avançado — PDTA/FPTI-BR — pelo auxílio por meio da linha de financiamento de bolsas.

Referências

[1] U. M. Fayyad, G. Platestsky-Shapiro e P. Smyth. From data mining to knowledge discovery: an overview. *American Association for Artificial Intelligence*, pág. 1–30, 1996.

[2] D. F. Honorato, E. A. Cherman, H. D. Lee, M. C. Monard e F. C. Wu. Construção de uma representação atributo-valor para extração de conhecimento a partir de informações semi-estruturadas de laudos médicos. In *Conferencia Latinoamericana de Informática*, pág. 1–12, Costa Rica, 2007.

[3] H. D. Lee. *Seleção de atributos importantes para a extração de conhecimento de bases de dados*. Tese de doutorado, Universidade de São Paulo, São Carlos, SP, 2005.

[4] J. J. R. Rocha. *Coloproctologia: princípios e práticas*. Atheneu, São Paulo, SP, 2005.

[5] J. C. M. Santos. Doença de crohn: aspectos clínicos e diagnósticos. *Revista Brasileira de Coloproctologia*, 19:276–285, 1999.

[6] D. F. Honorato, H. D. Lee, M. C. Monard, F. C. Wu, R. B. Machado, A. P. Neto e C. A. Ferrero. Uma metodologia para auxiliar no processo de construção de base de dados estruturadas a partir de laudos médicos. In *Encontro Nacional de Inteligência Artificial*, pág. 1–10, São Leopoldo, RS, 2005.

[7] E. A. Cherman, H. D. Lee, D. F. Honorato, J. J. Fagundes, J. R. N. Góes, C. S. R. Coy e F. C. Wu. Metodologia de mapeamento de laudos médicos para bases de dados: Aplicação em laudos colonoscópicos. In *II Congresso da Academia Trinacional de Ciências*, pág. 1–9, Foz do Iguaçu, PR, 2007.

[8] A. G. Maletzke, H. D. Lee, F. C. Wu, E. T. Matsubara, C. S. R. Coy, J. S. Fagundes e J. R. N. Góes. Uma metodologia para auxiliar no processo de mapeamento de formulários médicos para bases de dados estruturadas. In *X Congresso Brasileiro de Informática em Saúde*, Florianópolis, SC, 2006.

[9] A. G. Maletzke, H. D. Lee, W. Zalewski, R. F. Voltolini, E. T. Matsubara, C. S. R. Coy, J. J. Fagundes, J. R. N. Góes e F. C. Wu. Mapeamento de informações médicas

descritas em formulários para bases de dados estruturadas. In *VII Workshop de Informática Médica*, pág. 1–10, Porto de Galinhas, PE, 2007.

[10] W. Zalewski, H. D. Lee, F. C. Wu, A. G. Maletzke, C. S. R. Coy, J. J. Fagundes e J. R. N. Góes. Um sistema para construção automática de formulários e de uma base de dados relacionados à doença de crohn. In *XV Encontro Anual de Iniciação Científica*, pág. 1–3, Ponta Grossa, PR, 2006.

[11] F. Solimanpour, J. Sadri e C. Y. Suen. Standard databases for recognition of handwritten digits, numerical strings, legal amounts, letters and dates in farsi language. La Baule, France, 2006.

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

[13] O. D. Trier, A. K. Jain e T. Taxt. Feature extraction methods for character recognition - a survey. *Pattern Recognition*, 29:641–662, 1996.

[14] W. Zalewski, H. D. Lee, R. F. Voltolini, A. G. Maletzke, J. J. Fagundes, J. R. N Góes, C. S. R. Coy e F. C. Wu. Reconhecimento de caracteres manuscritos para o mapeamento de formulários médicos para bases de dados estruturadas. In *II Congresso da Academia Trinacional de Ciências*, pág. 1–10, Foz do Iguaçu, PR, 2007.

[15] S. Haykin. *Neural Networks - A Compreensive Foundation*. Prentice-Hall, USA, 2 ed., 1999.

[16] S. O. Rezende. *Sistemas inteligentes: fundamentos e aplicações*. Malone, São Paulo, SP, 2003.