Transferência de Formulários de Informações Médicas para Bases de Dados Estruturadas: Estudo de Caso

André G. Maletzke[1,3], Huei D. Lee[1], Willian Zalewski[1], Edson T. Matsubara[3], Richardson F. Voltolini[1], Cláudio S. R. Coy[2], João J. Fagundes[2], Juvenal R. N. Góes[2], Feng C. Wu[1,2]

[1]*Centro de Engenharias e Ciências Exatas – Universidade do Oeste do Paraná*
Laboratório de Bioinformática – LABI
Parque Tecnológico Itaipu – PTI
Foz do Iguaçu, Paraná, Brasil

[2]*Faculdade de Ciências Médicas – Universidade Estadual de Campinas*
Serviço de Coloproctologia
Campinas, São Paulo, Brasil

[3]*Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo*
Laboratório de Inteligência Computacional – LABIC
São Carlos, São Paulo, Brasil

Resumo
O crescente aumento no armazenamento de dados tem promovido a procura por métodos que auxiliem na análise e compreensão desse grande volume de dados. Nesse contexto, métodos computacionais, como Knowledge Discovery from Data Bases, vêm sendo amplamente aplicados. No entanto, é freqüente o registro de informações, manualmente, por meio de formulários impressos, impossibilitando a aplicação direta desses métodos. Neste trabalho é apresentada uma metodologia para o mapeamento de formulários médicos para uma representação em formato digital, de modo que essas informações possam ser analisadas computacionalmente. É também apresentado um estudo de caso considerando informações referentes à doença de Crohn.

Palavras – chave
Knowledge Discovery from Databases, Doença de Crohn, Inteligência Artificial.

1. Introdução

A evolução tecnológica tem promovido um aumento no volume e na variedade de informações armazenadas, cotidianamente, nas mais variadas áreas. Conseqüentemente, a análise dessas informações pelo ser humano, aplicando técnicas manuais, tornou-se complexa, demorada e propensa a falhas. Nesse sentido, a comunidade científica tem direcionado esforços no estudo e desenvolvimento de métodos e técnicas computacionais para tornar essa análise uma tarefa automática e menos complexa, apoiada pelo uso do computador. Atualmente essa área encontra-se em crescente evidência, sendo que um dos métodos de apoio à análise de grande volume de dados é o processo denominado *Knowledge Discovery from Databases* — KDD [1, 2]. Esse processo pode ser aplicado para a construção de modelos abstratos que representem algum conhecimento implícito existente nos dados analisados, os quais poderiam não ser identificados, de modo trivial, pelo ser

humano. O KDD é um processo iterativo e interativo, que está dividido em três fases: Pré-processamento de Dados, Mineração de Dados e Pós-processamento de Dados.

O Pré-processamento de dados é a fase que apresenta maior custo de tempo, cerca de 80% de todo o processo [3, 4]. Esta fase possui, basicamente, dois objetivos principais: conhecer o domínio da aplicação e dos dados e preparar os dados para que possam ser utilizados nas próximas fases. Algumas operações como preparação, limpeza e transformação de dados e atributos são realizadas nesta fase. Portanto, após a fase de Pré-processamento, é necessário que os dados estejam representados em um formato apropriado, geralmente sendo utilizado o formato atributo-valor [5]. Assim, considera-se esta fase como sendo uma das mais trabalhosas e demoradas de todo o processo, a qual é de fundamental importância dentro do KDD, pois é responsável por assegurar a qualidade dos dados que serão utilizados nas próximas fases. A fase de Mineração de Dados tem como finalidade a aplicação de métodos computacionais que permitem a extração e a representação do conhecimento existente nos dados. Assim, esta fase é realizada de modo iterativo, possibilitando que os parâmetros dos métodos aplicados sejam ajustados visando obter melhores resultados nos modelos construídos [2]. Após a fase de Pós-processamento, é necessário analisar esses modelos com o objetivo de avaliar o conhecimento extraído e interpretá-lo, juntamente com os especialistas da área, para que esse conhecimento seja validado e disponibilizado para aplicação.

Como mencionado, para que o processo de KDD possa ser aplicado é necessário que os dados estejam em um formato apropriado. Na área de medicina, os dados encontram-se freqüentemente em formatos não estruturados ou semi-estruturados, como laudos e formulários médicos, os quais podem conter informações referentes a, por exemplo, histórico e sintomas do paciente e hábitos alimentares. Diversos são os fatores relacionados à indisponibilidade desses dados em formatos adequados para a aplicação de processos como o KDD, entre os quais, a existência de ambulatórios médicos não informatizados, a necessidade de manterem-se registros impressos e por escolha de muitos especialistas da área de saúde que consideram que o contato com o paciente se torna mais pessoal [6].

Nesse contexto, surge a necessidade de desenvolver métodos e ferramentas que auxiliem, de maneira semi-automática, no processo de estruturação da informação, promovendo a padronização da informação e a redução no custo de tempo. Este trabalho é parte de um projeto maior denominado Análise Inteligente de Dados Aplicada ao Mapeamento de Dados Médicos — AIDMD — [5, 6, 7], e tem por objetivo apresentar uma metodologia que auxilia no mapeamento de formulários médicos para uma Base de Dados — BD — estruturada. A partir dessa metodologia foi desenvolvida uma ferramenta computacional, por meio da qual foi realizado um estudo de caso aplicado ao domínio médico relacionado à doença de Crohn.

Este trabalho está organizado da seguinte maneira: na Seção 2 é descrita a metodologia proposta, na Seção 3 é apresentado um estudo de caso aplicado a essa metodologia, na Seção 4 são apresentados os resultados e a discussão e na Seção 5 são apresentadas as conclusões e trabalhos futuros.

2. Metodologia Proposta

A metodologia proposta apresenta como objetivo auxiliar na coleta e no armazenamento de dados, de maneira estruturada em uma BD, a partir de formulários médicos de múltipla escolha. Essa metodologia está dividida em três etapas — Figura 1:

1. Geração de Formulários e Construção da Base de Dados;
2. Construção de Padrões sobre Formulários;
3. Mapeamento de Formulários e Preenchimento da Base de Dados.

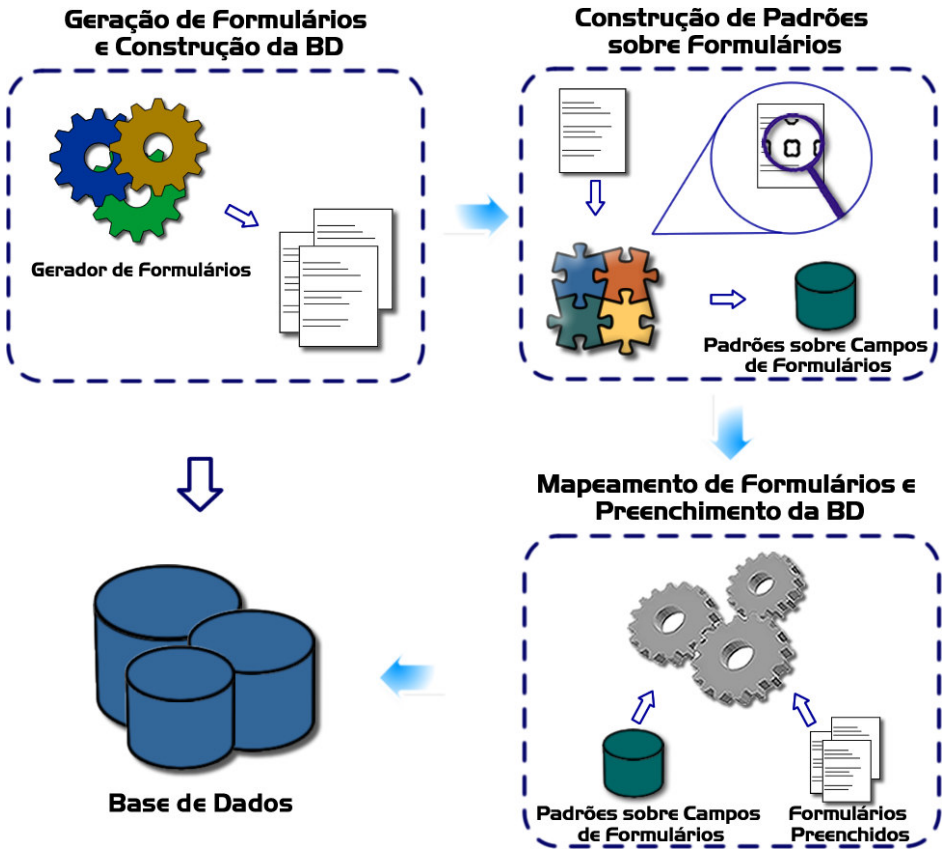

Figura 1 – Etapas da metodologia proposta.

Etapa 1: Geração de Formulários e Construção da Base de Dados

Esta etapa é responsável pela construção do formulário a partir de um conjunto de informações, de modo que essas informações sejam dispostas no formato de perguntas e respostas. É construída uma Base Dados para a qual essas informações serão mapeadas, futuramente, de acordo com as repostas selecionadas de cada pergunta. Para que essa BD seja preenchida é necessário que exista um conjunto de regras que indique a forma de preenchimento. Para tanto, nesta etapa é construído um Arquivo de Interpretação — AI — na linguagem *Extensible Markup Language* — XML[1], o qual é responsável por indicar o valor que deve ser armazenado na BD de acordo com as respostas que foram marcadas no formulário e o tipo de dado do atributo.

[1] http://www.w3.org/XML

Como mencionado, o formulário construído está estruturado no formato de perguntas e respostas, sendo que uma determinada pergunta poderá conter várias respostas, as quais são representadas pelo seu significado seguido de uma marca ou campo. Neste trabalho esses campos ou marcas são representados pela figura geométrica de um quadrado, identificando a região que deverá ser preenchida. Cada pergunta representa um atributo na BD e as respostas um dos possíveis valores que serão conferidos a esse atributo. Após a definição das perguntas e suas respectivas respostas, a construção do formulário é realizada de maneira automática por meio da geração de um arquivo na linguagem LATEX[2], o qual é processado pelo aplicativo PDFLaTeX, presente no conjunto de ferramentas disponibilizadas pelo MikTeX[3], gerando o formulário no formato *Portable Document Format* — PDF. Desse modo, é possível realizar a padronização em relação à formatação e minimizar o custo de tempo na construção do formulário, do AI e da BD. Na Figura 2 é ilustrada uma representação parcial de um formulário construído com informações referentes à sintomatologia de paciente.

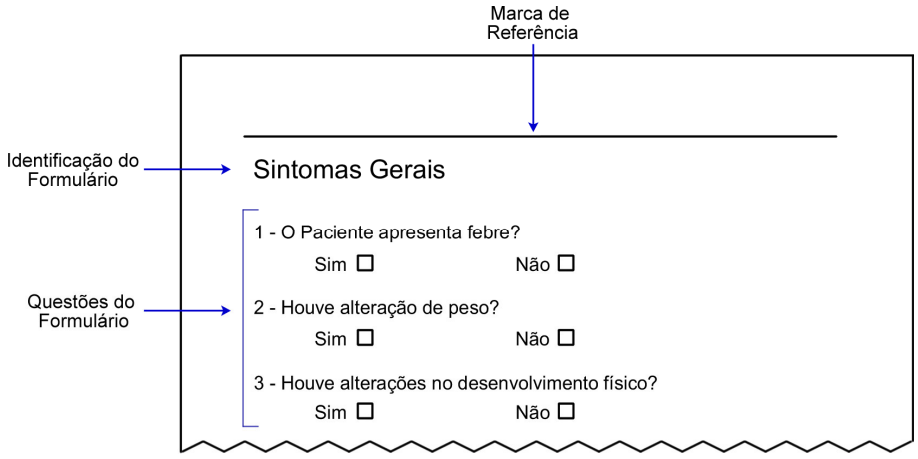

Figura 2 – Exemplo de um formulário.

Anteriormente à aplicação das Etapas 2 e 3 é necessário que os formulários sejam submetidos a uma fase de Pré-processamento, na qual esses são tratados para que possíveis ruídos e imperfeições indesejadas sejam removidos. Nessa fase são realizadas as seguintes operações: binarização, correção da escala e da inclinação e segmentação do formulário. A binarização da imagem de um formulário consiste em transformar uma imagem, representada por diversos níveis de cinza, em outra contendo somente duas cores: preto ou branco [8]. Após, o ângulo de inclinação e a escala do formulário são corrigidos, caso seja necessário, utilizando um método de interpolação. Neste trabalho foi utilizado o método de interpolação *Nearest Neighbor*, sendo determinado como padrão por meio da análise dos experimentos apresentados em [6]. Para isso, é realizada a leitura da Marca de Referência — MR — do formulário, representada por uma reta situada no cabeçalho do formulário e que se estende por toda sua largura, a qual é ilustrada na Figura 2. Posteriormente, realiza-se o processo de segmentação do formulário, baseado na localização e comprimento da MR, o qual consiste em dividir uma imagem em outras menores ou em extrair um

[2] http://www.latex-project.org
[3] http://www.miktex.org

fragmento da imagem original [9]. Assim, o formulário segmentado é formado pela área retangular, delimitada pelas coordenadas (x_0, y_0, x_f, y_f), onde x_0 e y_0 são referentes às coordenadas iniciais da MR e x_f e y_f são referentes aos comprimentos da MR e do formulário, respectivamente.

Etapa 2: Construção de Padrões sobre Formulários

Nesta etapa é realizada a construção dos padrões de um formulário por meio da identificação e armazenamento de informações como tamanho, localização e formato dos campos do formulário, as quais serão consideradas na etapa seguinte: Mapeamento de Formulários e Preenchimento da BD. Para tanto, é utilizada uma técnica baseada no algoritmo descrito em [10], a qual se baseia na análise das estruturas horizontais e verticais de um documento digitalizado. Desse modo, esta etapa está dividida em três fases:

1. Fase de Segmentação Horizontal;
2. Fase de Segmentação Vertical;
3. Fase de Construção de Padrões.

Fase de Segmentação Horizontal

Nesta fase a estrutura horizontal do formulário é analisada, por meio da geração de um Histograma de Projeção Horizontal — HH [10]. Esse processo consiste em contabilizar a quantidade de pixels pretos presentes em cada linha do formulário, sendo a espessura da linha equivalente a um pixel, permitindo construir um vetor de freqüência de pixels pretos para cada uma dessas linhas. Desse modo, a partir do vetor HH, o formulário é segmentado por meio da determinação de um Limiar de Segmentação — LS — em Banda de Texto — BT — e Banda Branca — BB. Neste trabalho é utilizado como padrão um limiar igual a um, o qual foi definido de acordo com as características de ruído do formulário. Uma BT consiste em um conjunto de linhas contínuas, na qual cada linha possui uma freqüência de pixels pretos, maior ou igual ao LS, enquanto uma BB é um conjunto de linhas contínuas com freqüência inferior ao LS. Conseqüentemente, na próxima fase é utilizado somente o conjunto de BTs, pois o conjunto de BBs representam regiões do formulário que não contém nenhuma informação relevante para análise.

Fase de Segmentação Vertical

Como mencionado, nesta fase serão consideradas somente as BTs, as quais serão analisadas por meio da geração de um Histograma de Projeção Vertical — HV. O HV consiste em um vetor de freqüências de cada coluna existente em uma BT, considerando a espessura das colunas equivalentes a um pixel. Posteriormente, cada HV gerado é particionado, por meio do LS, em um conjunto de HV menores, os quais irão compor um Conjunto de Busca — CB. Desse modo, cada elemento do CB é denominado de HVB, sendo que cada HVB pode representar uma marca, um ou mais caracteres e possíveis ruídos que não foram tratados na fase de Pré-processamento.

Por último, é realizada a construção da Base de Exemplos — BE. Para tanto, deve-se selecionar um subconjunto de marcas ou campos presentes no formulário, para que essas possam ser utilizadas, posteriormente, para rotular os elementos do CB. Um exemplo de uma marca de interesse é representado por um Histograma de Projeção Vertical da Base de Exemplos, denominado HVE. Assim, para cada marca selecionada é armazenado o seu HVE juntamente com suas respectivas coordenadas na BE.

Fase de Construção de Padrões

Nesta fase o objetivo consiste na construção de uma Base de Padrões — BP —, a qual é imprescindível para que o processo de mapeamento seja realizado. A BP contém informações referentes à localização de cada marca de interesse existente no formulário e é utilizada na próxima etapa para identificar quais marcas foram assinaladas em um formulário já preenchido. Para a construção da BP é realizada a classificação de todos os elementos existentes no CB considerando os elementos da BE. Para isso, foi implementado um classificador baseado no algoritmo *Nearest Neighbor* [11]. Esse método classifica um exemplo não rotulado por meio da atribuição da classe do seu vizinho mais próximo considerando uma medida de similaridade. O algoritmo implementado compara cada elemento do CB com os elementos existentes na BE utilizando uma técnica baseada na comparação de Séries Temporais, denominada *Dynamic Time Warping* — DTW [12]. Na Figura 3 é apresentada uma ilustração da comparação de exemplos da BE e do CB. Caso seja encontrada a similaridade entre o HVB em análise com algum HVE, a comparação é finalizada. Caso contrário, a comparação é realizada novamente, porém entre as freqüências H_x do HVE e as freqüências H_{x+t} do HVB, com x variando de 0 até o comprimento do HVE, t variando de 1 até n em intervalos de uma unidade e n sendo um valor inteiro menor ou igual à diferença de tamanho existente entre o HVE e o HVB. A medida de similaridade utilizada para determinar se um HVB é similar a um HVE foi a distância *Chi-Square* — χ^2 — [13, 14], a qual é amplamente utilizada na literatura na comparação de vetores de freqüência.

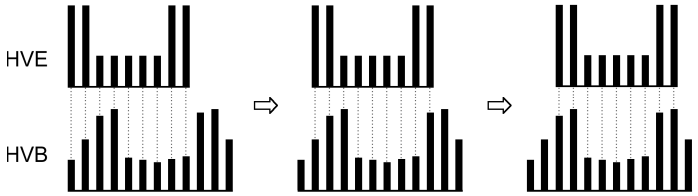

Figura 3 – Ilustrações de comparação de exemplos da BE e do CB.

Etapa 3: Mapeamento de Formulários e Preenchimento da Base de Dados

Esta etapa tem por objetivo mapear um formulário preenchido para uma Base de Dados, utilizando a BD e o Arquivo de Interpretação construídos na Etapa 1 e a Base de Padrões construída na Fase 3 da Etapa 2 da metodologia proposta. Desse modo, o formulário preenchido é submetido, primeiramente à fase de Pré-processamento para a correção de imperfeições, remoção de ruídos e segmentação da área de interesse. Posteriormente, são identificadas as marcas preenchidas no formulário, por meio da aplicação dos seguintes passos:

1. Segmentação de cada marca do formulário por meio das coordenadas contidas na Base de Padrões;
2. Contabilização da percentagem de pixels pretos que cada marca segmentada ocupa;
3. Identificação das alternativas marcadas por meio da determinação de um Limiar de Preenchimento — LP.

Após, o processo de preenchimento da BD inicia-se por meio da análise do AI, o qual irá indicar quais atributos da BD devem ser preenchidos, o valor e o tipo de dado desse atributo. Na Figura 4 são apresentados todos os passos realizados desde a utilização do

gerador de formulários para a construção do formulário padrão (a) até o mapeamento das informações contidas nos formulários preenchidos para a BD (b).

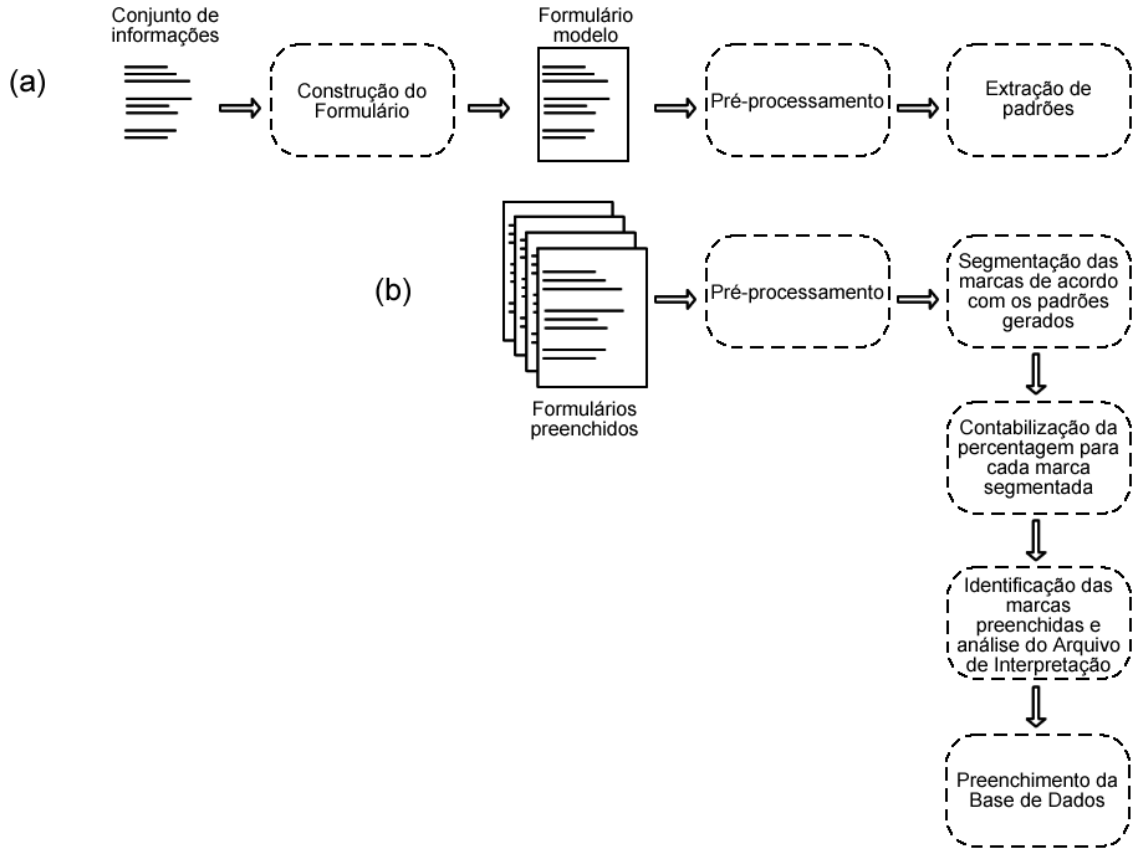

Figura 4 – Detalhamento das etapas da metodologia.

A partir da metodologia proposta foi desenvolvido um Sistema Computacional que provê todos os recursos apresentados anteriormente e atua como uma ferramenta para aplicação desta metodologia, proporcionando um ambiente no qual o usuário pode realizar ajustes nos diversos parâmetros existentes nos algoritmos. A construção do sistema foi realizada na linguagem de programação JAVA[4], utilizando o Sistema de Gerenciamento de Banco de Dados MySQL[5] e a API *Java Advanced Imaging* — JAI — para a manipulação dos formulários digitalizados [15].

3. Um Estudo de Caso Aplicado às Doenças Inflamatórias Intestinais

Um dos temas que tem despertado grande interesse pela comunidade científica de medicina refere-se às doenças inflamatórias intestinais, como a doença de Crohn. Essa enfermidade foi descrita pelo médico Burril B. Crohn e tem como principal característica afetar o sistema imunológico causando distúrbios em seu funcionamento [16, 17, 18]. Neste estudo de caso é apresentada a aplicação desta metodologia a formulários que contém informações sobre a doença de Crohn. Nesse contexto, o Laboratório de Bioinformática em parceria

[4] http://www.sun.com/
[5] http://www.mysql.com/

com o Serviço de Coloproctologia da Faculdade de Ciências Médicas da Universidade de Campinas elaboraram um protocolo contendo um conjunto variado de informações relacionadas à essa enfermidade, as quais foram consideradas relevantes para a análise. Dentre essas informações estão resultados clínicos de pacientes, sintomatologia, histórico alimentar entre outros. Neste trabalho um subconjunto dessas informações foi selecionado, juntamente com especialistas da área, para a aplicação da metodologia proposta. Esse subconjunto foi organizado em cinco questões e um total de 17 campos, a serem considerados para preenchimento, por formulário — Figura 5.

```
Informações referentes ao Estômago, Intestino Delgado, Cólon, Reto e Ânus
1. Mal estar pós-prandial: (sim/não) (tempo/duração)
    Duração: 0-5;5-15;15-30;30-60;>60
2. Náusea: (sim/não) (freqüência)
    Freqüência:1-2;2-3;>3
3. Vômito: (sim/não) (freqüência) (horário) (aspecto/cor)
    Freqüência: 1-2;2-3;>3
    Horário: Antes das refeições;Durante;Após as refeições
    Aspecto/cor: Esbranquiçada;Avermelhada;Amarelada;Escura
4. Soluço: (sim/não) (freqüência) (horário) (características) (duração)
    Freqüência:1-2;2-3;>3
    Horário: Antes das refeições;Durante;Após as refeições
    Características: Isolado/Crises
    Duração: 0-5;5-15;15-30;30-60;>60
5. Dispepsia: (sim/não) (freqüência) (intolerância a alimentos)
    Freqüência: 1-2;2-3;>3
    Intolerância a alimentos: (sim/não)
    Tipo de alimento: Gordura;Fibras;Apimentado;Outros
```

Figura 5 – Informações utilizadas para construir o formulário.

Desse modo, foi construído o formulário modelo, a Base Dados e o Arquivo de Interpretação, por meio da Etapa 1 e, posteriormente, foram identificados os padrões desse formulário com a aplicação da Etapa 2 da metodologia — Figura 4 (a). Após, cem cópias do formulário modelo foram geradas e distribuídas a dez colaboradores, os quais preencheram os formulários recebidos de modo "*ad libitum*", apenas fixando-se como critérios de preenchimento a realização do preenchimento próximo ao centro da marca e aplicando-se uma pressão normal de escrita. Uma vez que os formulários tenham sido preenchidos, os mesmos foram digitalizados e submetidos à fase de Pré-processamento e à Etapa 3 da metodologia como representado na Figura 4 (b). Como mencionado, esta última etapa é responsável pelo mapeamento do formulário para a BD por meio da análise do AI.

Neste estudo de caso, foi utilizado um LP igual a 10%, isto é, os campos cuja região segmentada apresentou uma percentagem de pixels pretos maior ou igual a 10% da área analisada foram considerados como preenchidos. Os formulários foram digitalizados fixando-se uma resolução de 85 *dots per inch* — DPI e uma configuração de cores em escala de cinza, considerando 256 tons de cinza.

4. Resultados e Discussão

Os resultados desse estudo de caso são apresentados na Tabela 1, na qual são mostradas as percentagens de mapeamento correto das informações para a BD e as respectivas quantidades de formulários.

Tabela 1 – Resultados do estudo de caso.

Número de Formulários	Percentagem de Mapeamento
87	100
8	94,12
3	88,24
2	82,35

Realizando-se uma análise mais detalhada sobre cada formulário mapeado observou-se que 87 formulários apresentaram 100,00% de suas marcas preenchidas mapeadas corretamente para a BD, ou seja, todas as 17 alternativas de cada um desses formulários foram mapeadas corretamente para a BD. Outros oito formulários apresentaram um mapeamento correto de 94,12%, três formulários obtiveram 88,24% e dois formulários apresentaram 82,35% das marcas preenchidas mapeadas corretamente para a BD. Desse modo, obteve-se uma precisão de 98,82% com desvio padrão 3,45%.

Deve-se observar também que, como mencionado, cada um dos 100 formulários apresenta 17 campos a serem mapeados para a Base de Dados, perfazendo um total de 1700 campos neste estudo de caso, dos quais 1680 foram mapeados corretamente para a BD.

Durante o processo de mapeamento dos formulários observou-se que dos 100 formulários preenchidos 12 apresentaram problemas na identificação da Marca de Referência, comprometendo seu mapeamento automático para a Base de Dados. Portanto, esses formulários foram pré-processados novamente com o ajuste da MR. Posteriormente, esses 12 formulários foram submetidos à Etapa 3 para o mapeamento para a BD

É importante ressaltar que os formulários que apresentaram uma percentagem baixa de mapeamento devem-se ao fato de que o preenchimento foi realizado próximo às bordas e utilizando-se de um marcador fino e/ou de cor clara. Esse tipo de marcação dificulta o reconhecimento, pois durante o processo de binarização os pixels de marcação com tons de cores mais claras podem ser confundidos com pixels do fundo. Outro fator importante que influencia diretamente no desempenho da metodologia está relacionado à identificação da MR, a qual é de suma importância para um correto mapeamento do formulário. Portanto, ruídos que tenham ocorrido, seja pelo manuseio dos formulários ou por influência dos dispositivos de reprodução ou digitalização dos formulários e que não puderam ser removidos na fase de Pré-processamento são atributos que influenciam de modo negativo na tarefa de mapeamento.

Como mencionado, fatores como espessura do marcador e cor são relevantes para obter-se um melhor desempenho no processo de mapeamento. Neste estudo de caso nenhum desses fatores foram fixados. Desse modo, a utilização de marcadores escuros e mais espessos poderá implicar em uma melhora significativa no desempenho do mapeamento.

5. Conclusão

A metodologia apresentada neste trabalho tem por objetivo a semi-automatização do processo de coleta de informações médicas, por meio da utilização de formulários de múltipla escolha para uma Base de Dados estruturada. Um estudo de caso também foi apresentado aplicado à doença inflamatória de Crohn. Nesse estudo de caso foi utilizado um subconjunto de informações consideradas relevantes a essa doença, o qual foi definido por especialistas.

A enfermidade de Crohn cada vez mais tem preocupado a médicos e pesquisadores, pois suas causas e cura ainda são pouco conhecidas e sua incidência na população mundial está aumentando nos últimos anos.

A metodologia apresentada auxiliou no processo de mapeamento de informações, tornando o processo uma tarefa menos custosa em relação ao tempo de mapeamento, e menos propensa a erros se comparada ao mapeamento manual. Além disso, torna possível a utilização de documentos impressos no registro das informações, característica considerada importante tanto para médicos especialistas na área médica quanto para estabelecimentos que necessitam manter registros na forma de documentos em papel ou até mesmo por não terem condições de serem informatizados.

Os especialistas avaliaram que a metodologia atendeu aos propósitos deste trabalho possibilitando a coleta de dados dispostos em documentos impressos para uma representação computacional. O sistema computacional desenvolvido a partir da metodologia será de grande utilidade em todas as etapas do mapeamento, pois é possível observar por meio dos resultados apresentados, que um dos fatores para obter-se um desempenho ainda melhor é a possibilidade de realizar ajustes nos parâmetros dos algoritmos de acordo com as características de um determinado formulário.

Trabalhos futuros incluem a aplicação de métodos como o KDD para a extração de possíveis padrões existentes nos dados coletados por meio desta metodologia. Outro trabalho inclui o desenvolvimento de um módulo para o reconhecimento e o mapeamento de caracteres manuscritos. Desse modo, espera-se um maior detalhamento dos dados coletados e conseqüentemente um enriquecimento dos padrões construídos a partir desses dados.

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

Referências

[1] S. O. Rezende, *Sistemas inteligentes: fundamentos e aplicações*, Manole, Barueri, Brasil, 2003.

[2] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth, "Knowledge discovery and data mining: towards a unifying framework", *Second International Conference on Knowledge Discovery and Data Mining*, p. 82–88, Menlo Park, USA, 1996.

[3] D. Pyle, *Data preparation for data mining*, Morgan Kaufmann, Califórnia, USA, 1999.

[4] I. H. Witten e E. Frank, *Data mining: practical machine learning tools and techniques*, Elsevier, San Francisco, USA, 2 ed., 2005.

[5] 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, *V Encontro Nacional de Inteligência Artificial*, p. 593-601, Porto Alegre, Brasil, 2005.

[6] 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, *X Congresso Brasileiro de Informática em Saúde*, p. 1-10, Florianópolis, Brasil, 2006.

[7] H. D. Lee, Seleção de atributos importantes para a extração de conhecimento de bases de dados, Tese de Doutorado, ICMC - USP, São Carlos, Brasil, 2005.

[8] R. C. Gonzalez e R. E. Woods, *Digital image processing*, Pretince Hall, New Jersey, USA, 2 ed., 2002.

[9] Y. J. Zhang, "Evaluation and comparison of different segmentation algorithms", *Pattern Recognition Letters*, vol. 18, p. 963–974, New York, USA, 1997, Elsevier Science Inc.

[10] T. H. Minh e H. Bunke, Analysis and understanding of giro check forms, Relatório técnico, Unviversity of Berne, Switzerland, 1992.

[11] I. H. Witten e E. Frank, *Data mining: practical machine learning tools and techniques with java implementations*, Morgan Kaufmann, San Francisco, USA, 2000.

[12] P. A. Morettin e C. M. Toloi, *Análise de séries temporais*, Edgard Blücher, São Paulo, Brasil, 2 ed., 2006.

[13] Ulysses Doria, *Introdução à Bioestatística: para simples mortais*, Elsevier, São Paulo, Brasil, 1999.

[14] S. Belongie, J. Malik, e J. Puzicha, "Matching shapes", *Eighth IEEE International Conference on Computer Vision*, p. 454–461, Vancouver, Canada, 2001.

[15] L. H. Rodrigues, *Building imaging applications with java technology*, Addison Wesley, Boston, USA, 2001.

[16] F. A. Quilici, *Colonoscopia*, Lemos-Editorial, São Paulo, Brasil, 1994.

[17] S. L. Robbins, R. S. Cotran, V. Kumar e T. *Collins, Patologia Estrutural e Funcional*, Guanabara Koogan, 6 ed., 2000.

[18] J. C. M. Santos JR, "Doença de Crohn: Aspectos Clínicos e Diagnósticos", *Revista Brasileira de Coloproctologia*, 1999.

Dados de Contato:

André Gustavo Maletzke[1] – andregm@icmc.usp.br
Huei Diana Lee[2] – huei@unioeste.br
Willian Zalewski[3] – willzal@gmail.com
Edson Takashi Matsubara[4] – edsontm@gmail.com
Richardson Floriani Voltolini[5] – rfv82@yahoo.com.br
Cláudio Sady Rodrigues Coy[6] – ccoy@terra.com.br
João José Fagundes[7] – jjfagundes@mpcnet.com
Juvenal Ricardo Navarro Góes[8] – rgoes@mpcnet.com.br
Feng Chung Wu[9] – wufc@unioeste.br

[1,2,3,5,9]*Centro de Engenharias e Ciências Exatas – Universidade do Oeste do Paraná, Laboratório de Bioinformática – LABI, Parque Tecnológico Itaipu – PTI, Caixa Postal 39, 85856-970 – Foz do Iguaçu, Paraná, Brasil.*

[6,7,8,9]*Faculdade de Ciências Médicas – Universidade Estadual de Campinas, Serviço de Coloproctologia, Caixa Postal 6111, 13083 – 970 – Campinas, São Paulo, Brasil.*

[1,4]*Instituto de Ciências Matemáticas e de Computação – Universidade de São Paulo, Laboratório de Inteligência Computacional, Caixa Postal 668,13560 – 970, São Carlos, São Paulo, Brasil.*