UM SISTEMA PARA MAPEAMENTO DE FORMULÁRIOS MÉDICOS PARA BASES DE DADOS ESTRUTURADAS

ANDRÉ GUSTAVO MALETZKE (UNIOESTE) , HUEI DIANA LEE , WU FENG CHUNG , EDSON TAKASHI MATSUBARA , CLÁUDIO SADY RODRIGUES COY , JOÃO JOSÉ FAGUNDES , JUVENAL RICARDO NAVARRO GÓES .

andregustavom@hotmail.com - UNIOESTE

O processo de Descoberta de Conhecimento em Bases de Dados auxilia na análise e compreensão de grandes quantidades de dados. Para que esses dados possam ser analisados por esse processo é necessário que sejam mapeados para uma base de dados, normalmente, no formato atributo-valor. Neste trabalho é proposta uma metodologia para o desenvolvimento de um de sistema que auxilia de maneira semi-automática o mapeamento de dados contidos em formulários médicos de múltipla escolha, relacionados à doença de Crohn, para uma base de dados estruturada.

metodologia; descoberta de conhecimento em bases de dados; doença de crohn

Introdução

O avanço tecnológico ocorrido nos últimos anos tem possibilitado o acúmulo crescente de dados nas mais diversas áreas do conhecimento, inclusive na área médica. Desse modo, a realização da análise desses dados por meio de ferramentas usuais torna-se uma tarefa cada vez mais complexa. Diversos métodos computacionais foram propostos para auxiliar essa análise, entre os quais o processo de Descoberta de Conhecimento em Bases de Dados - DCBD - [1], o qual possibilita a extração de padrões existentes nos dados, de modo a auxiliar no processo de tomada de decisão. Para que o processo de DCBD possa ser aplicado, o conjunto de dados deve estar em um formato adequado, normalmente atributo-valor. Nesse contexto, um dos temas que tem despertado interesse entre os pesquisadores da área médica está relacionado às doenças inflamatórias intestinais, como a doença de Crohn. Ela representa um desafio para médicos e pesquisadores, principalmente por apresentar aumento em sua incidência e por sua causa e cura serem ainda pouco conhecidas [2]. Processos, como o de DCBD, podem ser úteis como ferramentas de apoio para pesquisas relacionadas a essa doença. Este trabalho está inserido dentro do projeto de Análise Inteligente de Dados aplicado ao Mapeamento de Dados Médicos - AIDMD - [3], e tem como objetivo apresentar a metodologia utilizada para a construção de um sistema que auxilia no mapeamento semi-automático, para bases de dados estruturadas, de informações contidas em formulários médicos de múltipla escolha.

Material e Métodos

Os formulários utilizados neste trabalho estão estruturados na forma de perguntas e respostas definidas com o auxilio de especialistas. Cada pergunta representa um atributo na BD e as respostas, representadas pelo seu significado seguido de uma marca (quadrado), um dos possíveis valores que serão atribuídos para esse atributo. Ainda nesses formulários são inseridas marcas de referências, as quais irão auxiliar no processo de mapeamento desses formulários. Atualmente é utilizada uma única marca de referência, representada por uma linha horizontal, situada no cabeçalho de cada formulário. Para que esses formulários possam ser tratados computacionalmente, devem ser digitalizados por meio de um digitalizador óptico, com o qual os formulários são transformados em um arquivo constituído por elementos de imagem, chamados *pixels*.

O processo de mapeamento das informações contidas nesses arquivos de imagem para a BD foi dividido, basicamente, em duas etapas: construção de padrões sobre formulários e mapeamento dos formulários e preenchimento da BD - Figura 1.

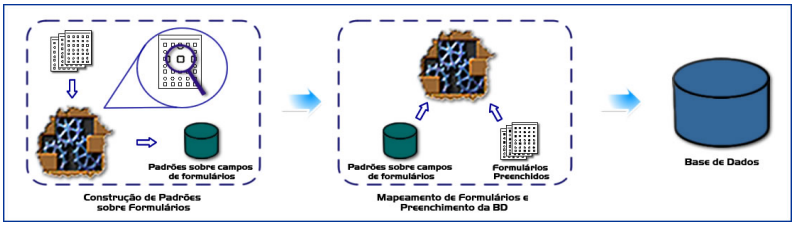

Figura 1 - Representação esquemática das etapas da metodologia proposta.

Antes de cada uma dessas etapas é necessário que esses arquivos de imagens passem por uma fase de preparação, denominada de pré-processamento. Nessa fase, são aplicados métodos de processamento de imagens para a binarização, ou seja, transformação dos valores dos *pixels* da imagem em 0 ou 1 e a segmentação da imagem de um formulário, na qual apenas a área de interesse delimitada pelas extremidades da linha de referência é mantida.

Com o término da fase de pré-processamento, é iniciada a primeira etapa da metodologia proposta, construção de padrões sobre formulários. O objetivo dessa etapa é mapear a localização das respostas (marcas) dos formulários não preenchidos e, armazenar essas informações, de maneira que possam ser utilizadas na etapa seguinte. O processo de mapeamento baseia-se no algoritmo descrito em [4], no qual a idéia é explorar o fato de que a grande parte dos documentos digitais possuem uma estrutura horizontal e vertical. Desse modo, é gerado um Histograma de projeção Horizontal - HH - do formulário, por meio da contabilização da quantidade de *pixels* pretos presentes em cada linha do formulário. Após a geração do HH, o formulário é segmentado por meio de um limiar de segmentação, em conjuntos de linhas contínuas, denominados Banda de Texto - BT -, as quais apresentam quantidade de *pixels* maior ao limiar estabelecido. Neste trabalho é utilizado um limiar igual a zero (0). Em um processo similar ao anterior, no qual é gerado o HH, é gerado o Histograma de projeção Vertical - HV - para cada BT. Cada HV, gerado, é segmentado por meio do limiar de segmentação, gerando um HV único para cada possível objeto do formulário, como marcas e caracteres, contido na região delimitada pela BT analisada. Os HVs de cada possível objeto são mantidos gerando um conjunto denominado de Conjunto de Busca - CB.

Ainda nessa etapa, o usuário seleciona por meio de uma interface gráfica um subconjunto de HV que corresponde às marcas de interesse, constituindo uma Base de Exemplos - BE. Essa BE é utilizada para reconhecer (mapear) as demais marcas de interesse no formulário. Para este propósito, foi implementado um algoritmo baseado no método de classificação *Nearest Neighbor* [5]. Esse algoritmo compara cada elemento do CB com os elementos contidos na BE. Para realizar essa comparação foi utilizada uma técnica baseada no método *Dynamic Time Warping* - DTW - [6], a qual pode ser aplicada na comparação de histogramas que possuem desenvolvimento semelhante, porém defasados um em relação ao outro. Para determinar a semelhança entre dois histogramas, do CB e da BE, foi utilizada pelo algoritmo uma medida de similaridade baseada na distância *Chi-Square*, a qual é amplamente aplicada na comparação de histogramas. Portanto, os HVs do CB que tiverem semelhança de acordo com algum elemento da BE, terão suas coordenadas mapeadas para uma base de dados denominada de Base de Padrões - BP.

A partir da construção da BP e dado um conjunto de formulários preenchidos, é iniciada a segunda etapa, mapeamento dos formulários preenchidos e preenchimento da BD, na qual são identificadas as respostas (marcas) que foram marcadas e, desse modo, a BD do sistema é preenchida. Para realizar a identificação das marcas preenchidas são realizados os seguintes passos:

1. Segmentação de cada marca contida no formulário, por meio das coordenadas armazenadas na BP;

2. Para cada marca segmentada é contabilizada a percentagem de *pixels* pretos que ocupam a área segmentada;

3. Caso a percentagem seja maior que a percentagem limiar aceita, então tal marca é considerada como marcada.

A partir da identificação das respostas marcadas no formulário, é analisado o Arquivo de Interpretação do preenchimento do formulário. Esse arquivo irá indicar qual atributo da BD deverá ser preenchido e qual será seu valor, de acordo com a resposta marcada.

Resultados e Discussão

A partir da metodologia proposta neste trabalho foi desenvolvido um sistema computacional, o qual tem como objetivo reconhecer e mapear, de maneira semi-automática, as informações contidas em formulários médicos, referentes à doença de Crohn. Os formulários utilizados neste trabalho foram construídos a partir de um Sistema Gerador de Formulários, o qual possui uma interface gráfica na qual o usuário fornece as perguntas e as respostas que irão compor o formulário. Durante o processo de digitalização desses formulários algumas imperfeições poderão ocorrer, como ruídos e inclinação do formulário dentro da área de digitalização. Com o intuído de reduzir as conseqüências da presença dessas imperfeições foram aplicados filtros para remoção de ruídos, e operações de transformação geométrica para a correção da inclinação, como a de rotação. Para a aplicação da operação de rotação foi utilizado o método de interpolação *Nearest Neighbor*, o qual é indicado para imagens nas quais não é desejado que novos valores de *pixels* sejam criados. O sistema foi desenvolvido na linguagem de programação JAVA, a qual utiliza o conceito de Orientação a Objeto - OO - e possui APIs para auxiliar na manipulação de imagens, entre as quais a API *Java Advanced Imaging* - JAI. A persistência dos dados é realizada por meio da linguagem *Structured Query Lenguage* - SQL - e o Sistema Gerenciador de Banco de Dados - SGBD - *mySql* está sendo utilizado para o armazenamento das informações, o qual é gratuito e oferece todos os recursos necessários para a realização deste trabalho.

Conclusões

O sistema construído permitiu que as informações contidas nos formulários médicos sejam mapeadas para a BD. Essa BD poderá, posteriormente, ser utilizada no processo de Descoberta de Conhecimento em Bases de Dados, com o intuito de extrair padrões desses dados, os quais poderão auxiliar aos especialistas nas pesquisas relacionadas à doença de Crohn. Como trabalho futuro propõe-se a incorporação a esta metodologia de um módulo para o reconhecimento de caracteres manuscritos, o qual possibilitará maior flexibilidade e precisão na coleta dos dados.

Agradecimentos

À Fundação Parque Tecnológico Itaipu - FPTI - pela concessão de bolsa de iniciação científica.

Referências Bibliográficas

1. U. Fayyad; G. Piatetsky-Shapiro; P. Smyth in Second International Conference on Knowledge Discovery and Data Mining. Menlo Park, CA, 1996, p. 82-88.

2. F. Cordeiro; J. S. M. Filho; J. C. Prolla. *Endoscopia Digestiva*. Ed.; MEDSI, Rio de Janeiro, 1994.

3. H. D. Lee. Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, Universidade de São Paulo; ICMC-USP, 2005.

4. T. H. Minh; H. Bunke. *Analysis and Understanding of GIRO Check Forms*. Relatório Técnico, Universidade de Berne, 1992.

5. T. M. Mitchell. *Machine learning*. Ed.; McGraw-Hill, Boston, 1997.

6. J. C. Felipe; A. J. M. Traina in IX Congresso Brasileiro de Informática em Saúde, 2004, p. 234-239.