República Federativa do Brasil

Ministério do Desenvolvimento, Indústria, Comércio e Serviços

Instituto Nacional da Propriedade Industrial

(11) PI 1005740-4 B1

(22) Data do Depósito: 01/10/2010

(45) Data de Concessão: 12/03/2024

(54) Título: MÉTODO PARA MAPEAMENTO DE DOCUMENTOS TEXTUAIS PARA BASES DE DADOS ESTRUTURAS UTILIZANDO ONTOLOGIAS

(51) Int.Cl.: G06F 19/00.

(73) Titular(es): UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP.

(72) Inventor(es): WU FENG CHUNG; CLÁUDIO SADDY RODRIGUES COY; HUEI DIANA LEE; JOÃO JOSÉ FAGUNDES; CARLOS ANDRÉS FERRERO; RENATO BOBSIN MACHADO; ANDRÉ GUSTAVO MALETZKE; WILLIAN ZALEWSKI; RAQUEL FRANCO LEAL; MARIA DE LOURDES SETSUKO AYRIZONO; LUIZ HENRIQUE DUTRA DA COSTA.

(57) Resumo: MÉTODO PARA MAPEAMENTO DE DOCUMENTOS TEXTUAIS PARA BASES DE DADOS ESTRUTURADAS UTILIZANDO ONTOLOGIAS Refere-se o presente pedido de patente de invenção a um método para mapeamento de dados textuais descritos em língua natural (arquivos digitais de texto como laudos e relatórios) para um formato estruturado (bases de dados computacionais (BD) com campos e valores que estes podem assumir, bem definidos).

"MÉTODO PARA MAPEAMENTO DE DOCUMENTOS TEXTUAIS PARA BASES DE DADOS ESTRUTURADAS UTILIZANDO ONTOLOGIAS"

Campo da Invenção

Refere-se o presente pedido de patente de invenção a um método para mapeamento de dados textuais descritos em língua natural (arquivos digitais de texto como laudos e relatórios) para um formato estruturado (bases de dados computacionais (BD) com campos e valores que estes podem assumir, bem definidos).

A aplicação industrial se baseia na possibilidade de se mapear dados contidos em textos digitais descritos, em língua natural, para bases de dados computacionais e após, analisar esses dados por meio de métodos descritivos e métodos analíticos convencionais e não-convencionais. Essa análise pode ser realizada, por exemplo, por meio de métodos para a extração de padrões, relacionados à área de Mineração de Dados e Inteligência Artificial como árvores de decisão, redes neurais, entre outros. Esses modelos podem ser utilizados para explorar conceitos embutidos nos dados, tarefas de predição e apoio a processos de tomada de decisão, desde a área científica até apoio a processos de gestão.

Vale ressaltar que o método proposto, objeto do presente pedido de patente, pode ser aplicado para qualquer área do conhecimento, não se restringindo à área médica, a qual foi o tema inicial de aplicação. Qualquer domínio no qual existam documentos digitais em texto pode se beneficiar da aplicação desse método, pois ao mapear os dados contidos nesses documentos para BD, esses dados poderão ser analisados e será possível, por exemplo, conhecer padrões de comportamento de equipamentos diante de certas condições de uso, comportamentos mercadológicos de clientes e padrões encontrados em diversos processos, incluindo a área de gestão, documentados através de documentos textuais.

Fundamentos da Invenção

Inicialmente este método foi proposto para aplicação na área médica, a qual apresenta uma grande quantidade de documentos textuais em formato digital, como laudos e relatórios. O objetivo era o mapeamento de informações contidas nesses documentos digitais textuais para um formato estruturado, como bases de dados estruturadas (BD) computacionais [tabela atributo-valor (TAV)]. Após esse mapeamento, os dados, que antes se encontravam em formato de texto, podem ser analisados por meio de diversos tipos de métodos computacionais para gerar desde estatística descritiva até modelos que representam padrões encontrados nos dados.

O método proposto pode também ser usado para mapear texto em formato digital, na forma de laudos, relatórios e similares, de qualquer área do conhecimento, para bases de dados computacionais estruturadas.

Até onde conhecemos o problema e diante das buscas de anterioridade realizadas, a tarefa relacionada ao mapeamento de informações, no nível como proposto por nós, é atualmente resolvido de forma manual, isto é, especialistas necessitam ler os documentos e transcrever manualmente as informações para a BD.

A abordagem do mapeamento manual apresenta duas principais desvantagens:

1. tempo despendido para o mapeamento; e
2. subjetividade na interpretação das informações.

A primeira desvantagem está relacionada ao tempo despendido para realizar a tarefa de ler os documentos e transcrever manualmente as informações, tarefa esta que necessita tanto da participação de especialistas da área referente ao conteúdo do documento como de especialistas da área computacional para verificar o formato e o modo como se realiza a transcrição das informações para a BD. A segunda desvantagem está relacionada à subjetividade e a diversificação da interpretação das informações contidas nos documentos, já que se mais de um especialista participa da leitura do documento e transcrição das informações, podem ocorrer diferenças, por

exemplo, entre o nível de detalhamento das informações, decorrentes da subjetividade na interpretação das informações e do processo de mapeamento.

Assim sendo, torna-se necessária a construção de um método que permita representar o conhecimento presente nos documentos e mapear as informações contidas nesses documentos de modo objetivo para BD. As ontologias apresentam flexibilidade e recursos necessários para representação de conhecimento e para utilização dessa representação por métodos computacionais com o objetivo de mapear informações descritas em língua natural para BD estruturadas.

A seguir são apresentadas algumas tecnologias relacionadas ao presente pedido de patente de invenção:

- (US 6292771 / US 09/164,048): codificação de conceitos para BD (utilizando códigos predefinidos como o ICD9). O objetivo deste invento não é o mesmo. A construção e o mapeamento das informações são baseados apenas nas informações contidas nos dados extraídos de modo estatístico. Se os dados forem não-representativos, há possibilidade de mapeamento incompleto ou errôneo das informações. No presente pedido de patente de invenção, a construção da ontologia é realizada baseada nas informações contidas nos documentos digitais, com base no CFU, *n-gramas*, mas principalmente com a supervisão de especialistas, característica que previne o mapeamento de informações incompletas/errôneas em uma construção totalmente automatizada da estrutura que será utilizada para mapeamento das informações para a BD;
- (US 09/370,329): síntese de informação e/ou recuperação de informação usando linguagens de marcação (XML) para rotular informações no próprio documento original e transformar para o padrão HL7. O objetivo principal desse invento é rotular as informações consideradas importantes no texto original para, principalmente, as tarefas de indexação e recuperação de documentos. Utilizam para isso alguma forma estruturada para armazenar essas informações como XML, que podem ser guardadas em uma BD para fins de busca (durante recuperação de documentos). Após, propõe também a

transformação dessas informações para o padrão HL7 (de codificação e transmissão de informações de pacientes). Não se utiliza ontologias para representar os conceitos. O objetivo final não é o mesmo; no presente pedido de patente de invenção, geramos uma BD geral (os especialistas determinam que informações serão mapeadas) com o objetivo de realizar análises e encontrar padrões inerentes aos dados, através da construção de modelos diversos, por exemplo, usando Mineração de Dados e técnicas de Inteligência Artificial. No presente pedido de patente de invenção ao se preencher a BD com os valores das informações encontradas nos documentos, deixa-se o usuário livre para, posteriormente, construir qualquer tipo de modelo que julgue interessante, como árvores de decisão, redes neurais, modelos estatísticos entre outros. Além disso, no presente pedido de patente de invenção, também pode se transformar os dados mapeados para HL7 ou qualquer outro padrão;

- (US 20040172237): síntese de informação e/ou recuperação de informação usando linguagens de marcação (Natural Markup Language (NML)) para criar objetos de programa para posterior execução. O objetivo deste invento é transformar uma sentença em inglês para um conjunto de objetos de software que são passados posteriormente a um programa para execução. Como exemplo de aplicação tem-se a possibilidade de associação de uma interface em Linguagem Natural a qualquer sistema. Os objetos do domínio são capturados usando NML e podem ser transcritos para uma BD. São utilizadas outras técnicas para realizar o mapeamento de dados não-estruturados para uma base de dados, como uma árvore para *parsing* e não ontologias como no presente pedido de patente de invenção. O objetivo final não é exatamente o mesmo de nossa proposta. Embora a construção de BD a partir de texto seja possível, o objetivo específico proposto está mais ligado a síntese de conteúdo e/ou recuperação de informação do que mapeamento de dados a partir de texto plano para BD com o intuito de realizar análises posteriores. A aplicação típica está relacionada à construção de BD baseada em atributos e parâmetros a partir de conteúdo ou produtos em *Websites*. Com adaptação, essa invenção poderia realizar tarefa semelhante ao nosso objetivo.

No entanto, no presente pedido de patente de invenção, ao preenchermos a BD com os valores das informações encontradas nos documentos, deixamos o usuário livre para, após, construir qualquer tipo de modelo que julgue interessante, para analisar os dados;

- (US 20060026203): construção de modelos usando modelos associativos após terem mapeados os dados para uma BD que contém meta conhecimento na forma de nome-verbo-nome ou conceito-relação-conceito. Uma parte do objetivo final deste invento é semelhante: construir uma base de dados (BD), a partir de texto. No entanto, nesse invento a BD contém meta conhecimento na forma de nome-verbo-nome ou conceito-relação-conceito e constrói após "conhecimento" na forma de modelos associativos. O objetivo final não é o mesmo do proposto no presente pedido de patente de invenção, no qual ao preenchermos a BD com os valores das informações encontradas nos documentos, deixamos o usuário livre para, após, construir qualquer tipo de modelo que julgue interessante, para analisar os dados;

- (JP 2006-178982): construção de modelos, a partir de ontologias, que contém padrões construídos a partir dos dados. O objetivo deste invento é encontrar padrões nos dados a partir de conceitos contidos em uma ontologia e não realizar o mapeamento de dados, a partir de documentos de texto. Desse modo, o objetivo final não é o mesmo do proposto no presente pedido de patente de invenção, no qual ao preenchermos a BD com os valores das informações encontradas nos documentos, deixamos o usuário livre para, após, construir qualquer tipo de modelo que julgue interessante, para analisar os dados;

- (JP 2003-233528): mapeamento entre esquemas de dados. O objetivo deste invento é mapear um esquema de dados em outro mapeando o primeiro esquema em uma ontologia e o segundo esquema em outra ontologia e derivando um modelo de conversão entre os dois e não realizar o mapeamento de dados a partir de documentos de texto. Desse modo, o objetivo final não é o mesmo do proposto no presente pedido de patente de

invenção, no qual ao preenchermos a BD com os valores das informações encontradas nos documentos, deixamos o usuário livre para, após, construir qualquer tipo de modelo que julgue interessante, para analisar os dados.

A essência da presente invenção está em permitir o mapeamento de informações (tão detalhadas quanto possível e segundo a determinação dos especialistas) existentes em documentos digitais de texto (sem estrutura/planos) para uma BD estruturada.

Para tanto, uma ontologia, construída com base em documentos digitais do domínio e supervisão de especialistas da área, é utilizada para o mapeamento dos conceitos contidos nos documentos a serem mapeados. A intensa participação de especialistas nessa tarefa garante a eficiência e a precisão do mapeamento, bem como a determinação do nível de detalhes que se deseja mapear, diferentemente da invenção (US 09/164,048), na qual o nível de mapeamento está restrito aos códigos pré-existentes do ICD9. Situação semelhante é a obtida pela invenção (US 09/370,329) em relação ao padrão HL7. A ontologia utilizada em nossa proposta pode ser facilmente expandida e atualizada com a utilização da lista de termos não-processados, não sendo necessário re-treinar um modelo de mapeamento, como no caso da invenção (US 09/164,048) e não sendo descartada como no caso (US 09/370,329).

Em nossa proposta, não há um tipo de modelo de padrões pré-definido, encontrados nas informações contidas nos textos, como nas invenções (US 20060026203) e (JP 2006-178982). O usuário, após utilizar o método para mapear as informações para uma BD estruturada, pode escolher desde realizar uma simples estatística descritiva, até aplicar métodos de extração de padrões, relacionados à área de Mineração de Dados e Inteligência Artificial como árvores de decisão, redes neurais, entre outros. Esses modelos podem ser utilizados para explorar conceitos embutidos nos dados, tarefas de predição e apoio a processos de tomada de decisão.

Nossa proposta também não está relacionada ao problema de conversão de um formato de dados em outro como tratam as invenções (US

20040172237) e (JP 2003-233528) quer seja usando linguagens de marcação ou ontologias para tanto.

Breve descrição da Invenção

Refere-se o presente pedido de patente de invenção a um método para mapeamento de dados textuais descritos em língua natural (arquivos digitais de texto como laudos e relatórios) para um formato estruturado (bases de dados computacionais (BD) com campos e valores que estes podem assumir, bem definidos). Tal método dá-se da seguinte forma:

FASE 1 (Etapa 1): Processo de identificação de padrões, composto, mas não restrito, pelas seguintes tarefas:

1.1.a) Identificação de frases únicas: identificação de todas as frases dos documentos e eliminadas as frases repetidas formando um Conjunto de Frases Únicas (CFU);

1.1.b) Redução de quantidade e de ambigüidade dos termos contidos no CFU para auxiliar na identificação dos padrões a serem mapeados através de:

i. Aplicação da técnica de *Stemming e/ou Lematização*: redução ao radical das palavras e/ou extração de inflexões;
ii. Definição da lista de *stopwords* (palavras a serem filtradas, removidas);
iii. Construção do Arquivo de Padronização contendo palavras ou expressões que serão substituídas por outras equivalentes a fim de unificar a descrição dos termos;
iv. Geração de *n-gramas* que constituem uma seqüência de n letras ou palavras, a fim de identificar as unidades terminológicas de maior frequência nos documentos.

Produtos/Artefatos gerados pela FASE 1 (Etapa 1): CFU original, lista de *stopwords*, arquivo de padronização e lista de *n-gramas*.

FASE 1 (Etapa 2): Processo de construção da estrutura para representação do conhecimento (ontologia) e da definição da estrutura da BD,

com o auxílio de especialistas, composto, mas não restrito, pelas seguintes tarefas:

1.2.a) Construção da ontologia, baseada no CFU e nos *n-gramas*, a qual é utilizada para representar os conceitos e os objetos do domínio de mapeamento dos documentos e dos atributos da BD (TAV);

1.2.b) Definição da TAV através da definição dos atributos que irão compor a TAV e os possíveis valores que podem ser preenchidos para cada atributo da tabela, com base em interações com especialistas e nos artefatos construídos na FASE 1 (Etapa 1) (CFU, arquivo de padronização, *n-gramas*).

Produtos/Artefatos gerados pela FASE 1 (Etapa 2): Ontologia específica para domínio de aplicação e BD (TAV) correspondente, que será preenchida pela FASE 2.

FASE 2: Processo de mapeamento das informações contidas nos documentos textuais para a BD estruturada (TAV), através, mas não restrito, das seguintes tarefas:

FASE 2 (Etapa 1): Padronização dos documentos originais, conforme estabelecido no arquivo de padronização e na lista de *stopwords* construídos no FASE 1 (Etapa 1);

FASE 2 (Etapa 2): Aplicação do Algoritmo de Mapeamento, que é organizado em três tarefas (executadas para cada documento digital de texto a ser mapeado):

2.2.a) Divisão do documento em um conjunto de sentenças;

2.2.b) Processamento das sentenças com a identificação dos termos e vinculação da respectiva propriedade, conforme sua classificação definida na ontologia; os termos não identificados são adicionados a uma lista de termos não-processados. Em seguida o conjunto de propriedades mapeadas é associado, segundo as regras de associação definidas, e são identificados quais atributos devem ser preenchidos, assim como os valores que serão preenchidos. Essa tarefa então é repetida em cada sentença do laudo até que todas tenham sido

processadas. Cria-se então um conjunto de atributos e seus respectivos valores para esse conjunto de sentenças;

2.2.c) Preenchimento de atributos não-mapeados através da consulta à ontologia para verificação dos atributos ausentes no conjunto de atributos mapeados, construído na tarefa anterior, e também para identificação do valor padrão a ser utilizado para preenchimento, conforme definido na ontologia, para cada um dos atributos que não foram mapeados.

Produtos/Artefatos gerados pela FASE 2: BD (TAV) preenchida com as informações contidas nos documentos digitais de texto e lista de termos não-processados, a qual pode ser utilizada para expandir, facilmente, a ontologia construída.

Breve descrição das figuras

A figura 1 demonstra um fluxograma do processo para mapeamento de dados textuais descritos em língua natural (arquivos digitais de texto como laudos e relatórios) para um formato estruturado (bases de dados computacionais (BD) com campos e valores que estes podem assumir, bem definidos).

A figura 2 demonstra um exemplo de tabela de documento digital de texto original antes do seu processamento (Trecho de um laudo de Endoscopia Digestiva Alta (EDA)).

A figura 3 demonstra um exemplo de tabela de documento digital de texto após parte do processamento, mais especificamente da padronização (Trecho de laudo de EDA após padronização).

A figura 4 refere-se a uma representação geral dos conceitos da ontologia construída na primeira iteração.

Descrição detalhada da invenção

Nas diversas áreas do conhecimento, é comum a existência de documentos digitais de texto, como laudos e relatórios, os quais são confeccionados por profissionais para descrever e registrar ações relacionadas a atividades de suas áreas. Por exemplo, na área médica, é comum encontrar

informações na forma de laudos médicos, que apresentam a descrição textual de um exame. Esses laudos normalmente são compostos por sentenças em que o médico descreve, em língua natural, as observações a respeito de um exame e/ou condições de saúde de um paciente. A proposta aqui descrita foi originalmente desenvolvida para resolver o problema na área médica, mais especificamente no tema de exames de Endoscopias Digestivas Alta e Baixa, os quais são exames freqüentemente descritos no formato de laudo textual. No entanto, essa proposta não se restringe à área médica e pode ser aplicada a todas as áreas do conhecimento para as quais existam documentos em forma de texto digital. Como exemplo para se explicar a invenção proposta, será usado o laudo de Endoscopia Digestiva Alta (EDA).

Para que as informações registradas nesses laudos possam ser analisadas, por meio de estatística descritiva, analítica ou para extração de padrões por meio de Mineração de Dados, por exemplo, é estritamente necessário que essas informações estejam representadas em um formato estruturado, como o da tabela atributo-valor (TAV). Nesse formato de dados, cada linha corresponde às informações de um laudo e cada coluna corresponde às características, variáveis, conceitos ou atributos que se deseja mapear dos laudos. As TAV podem ser consideradas uma representação abstrata de bases de dados (BD) estruturadas computacionais e não estão vinculadas a nenhuma tecnologia para manipulação de BD. Neste documento, utilizaremos indistintamente os termos TAV e BD.

O mapeamento manual, ou seja, realizado através da leitura, interpretação e transcrição, desses laudos em BD apresenta-se como um método lento, além de apresentar um determinado grau de subjetividade, pois pode ser influenciado por fatores subjetivos dos especialistas que realizam essa tarefa. Para reduzir o tempo e a subjetividade do mapeamento desses dados foi desenvolvido um método de mapeamento de informações contidas em documentos digitais em texto, por meio de Ontologias.

O método de mapeamento de documentos é organizado em duas fases:

FASE 1 (Identificação de Padrões e Construção da Estrutura de Representação do Conhecimento) - é composta por duas etapas, realizadas com o auxílio de especialistas:

FASE 1 (Etapa 1): Identificação de padrões, composta, mas não restrita, pelas seguintes tarefas:

1.1.a) Identificação de frases únicas: identificação de todas as frases dos documentos e eliminação das frases repetidas, formando um Conjunto de Frases Únicas (CFU);

1.1.b) Redução de quantidade e de ambigüidade dos termos contidos no CFU para auxiliar na identificação dos padrões a serem mapeados através, mas não restrita, de:

i. Aplicação da técnica de *Stemming e/ou Lematização*: as palavras com distintas morfologias de gênero, grau e número são reduzidas ao seu radical e/ou pode ser realizada a substituição de diferentes inflexões das palavras pelas formas canônicas correspondentes, como o infinitivo de um verbo ou o masculino singular de um substantivo;

ii. Definição da lista de *stopwords*: definição de uma lista de palavras que serão filtradas dos documentos para reduzir o tamanho do CFU. Essas palavras são definidas, em conjunto com especialistas, como sendo não-importantes para o mapeamento das informações, como artigos, conjunções e preposições;

iii. Construção do Arquivo de Padronização: construção de um arquivo de padronizações, no qual são definidas palavras ou expressões que serão substituídas por outras equivalentes a fim de unificar a descrição dos termos;

iv. Geração de *n-gramas*: geração de *n-gramas*, uma seqüência de n letras ou palavras, a fim de identificar as unidades terminológicas de maior frequência nos

documentos e que possivelmente estarão relacionadas às características a serem mapeadas.

FASE 1 (Etapa 2): Construção da estrutura para representação do conhecimento, com o auxílio de especialistas, composta pelas seguintes tarefas:

1.2.a) Construção da ontologia: construção da ontologia, baseada no CFU e nos *n-gramas*, a qual é utilizada para representar os conceitos e os objetos do domínio de mapeamento dos documentos e dos atributos da TAV. Em outras palavras, a ontologia representará os termos dos documentos, as relações entre si e com os atributos da TAV, os quais também são representados na ontologia;

1.2.b) Definição da TAV: definição dos atributos que irão compor a TAV e os possíveis valores que podem ser preenchidos para cada atributo da tabela, com base em interações com especialistas e com o CFU, o arquivo de padronização, a geração de *n-gramas* e aplicação de *stemming/lematização* da etapa anterior.

FASE 2 (Mapeamento dos Documentos para a Base de Dados Estruturada) – é composta por duas etapas:

FASE 2 (Etapa 1): Realização de padronização dos documentos originais a serem mapeados para a BD, conforme estabelecido no arquivo de padronização e na lista de *stopwords* construídos na FASE 1. A aplicação das padronizações acarreta na generalização de determinadas informações contidas nos laudos abstraindo parte do conteúdo;

FASE 2 (Etapa 2): Processamento dos documentos padronizados por um Algoritmo de Mapeamento que preenche, com base na ontologia construída, a BD (TAV) definida na primeira fase e também gera uma lista contendo todos os termos não-processados. Esse Algoritmo de Mapeamento é organizado em três tarefas que são executadas para cada documento:

2.2.a) Divisão do documento em sentenças: o documento é dividido em um conjunto de sentenças;

2.2.b) Processamento das sentenças: os termos são identificados e a cada termo é vinculada uma propriedade, conforme sua classificação definida na ontologia, e os termos não-identificados são adicionados a uma lista de termos não-processados. Em seguida, o conjunto de propriedades mapeadas é associado segundo as regras de associação definidas na ontologia e são identificados quais atributos devem ser preenchidos, assim como os valores que serão preenchidos. Essa etapa então é repetida em cada sentença do laudo até que todas tenham sido processadas. Cria-se então um conjunto de atributos e seus respectivos valores para esse conjunto de sentenças;

2.2.c) Preenchimento de atributos não-mapeados: consulta à ontologia para verificação dos atributos ausentes no conjunto de atributos mapeados, construído na etapa anterior, e também para identificação do valor padrão a ser utilizado para preenchimento, conforme definido na ontologia, para cada um dos atributos que não foram mapeados.

Após a execução dessas três tarefas para todos os laudos, obtêm-se uma TAV preenchida e uma lista de termos não-mapeados. Esse algoritmo apresenta complexidade linear em relação ao número total de termos a serem mapeados, isto é, o tempo de execução do algoritmo é linearmente dependente da quantidade de documentos a serem processados e da quantidade de termos a serem processados em cada documento.

A lista de termos não-processados pode ser utilizada para expandir, facilmente, a ontologia previamente construída.

A tabela abaixo demonstra uma visão geral do método, como um todo, para Mapeamento de Documentos Textuais para Bases de Dados Estruturadas, o qual é composto basicamente de duas etapas principais:

Fase 1 – identificação de padrões e construção da ontologia [com a definição da BD (TAV)];

Fase 2 – mapeamento das informações contidas nos documentos digitais textuais para a BD (TAV).

A comparação das figuras 2 e 3 demonstra a evolução dos resultados obtidos com o processamento do texto digital, objeto do presente pedido de patente de invenção.

A tabela abaixo demonstra uma ilustração da TAV, onde Ei denota os casos, documentos ou laudos (no exemplo descrito aqui), Xj os atributos ou características e xij os valores correspondentes ao atributo Xj para o caso Ei, para i = 1..N e j = 1..M

	Atributos			
Exemplos	X_1	X_2	...	X_M
E_1	x_{11}	x_{12}	...	x_{1M}
E_2	x_{21}	x_{22}	...	x_{2M}
E_3	x_{31}	x_{32}	...	x_{3M}
⋮	⋮	⋮	⋱	⋮
E_N	x_{N1}	x_{N2}	...	x_{NM}

Breve descrição sobre Ontologias

Os diversos domínios do conhecimento existentes registram um grande volume de dados, entretanto com esse aumento torna-se difícil a análise e a manutenção desses dados. De modo a prover suporte ao arranjo dos dados de forma organizada e melhorar a seleção, o processamento e a recuperação de dados pode-se fazer uso de diferentes técnicas, dentre elas estão inseridas as de organização e representação do conhecimento, as quais permitem, por exemplo, a construção de ontologias.

Na filosofia, a ontologia (grego onto, "ser" no sentido de existir, e logia, "ciência, estudo") trata da natureza do ser, da realidade e das questões metafísicas em geral (WordNet, 2006).

Na área da computação, uma ontologia é geralmente definida como uma definição explícita de uma conceitualização compartilhada, ou seja, é uma estrutura em que se define de modo explícito e formal os conceitos, as instâncias, as relações, as restrições e os axiomas correspondentes a um domínio do conhecimento seguindo uma terminologia comum ao domínio (Gruber, 1995; Carvalheira, 2007).

Como mencionado, na filosofia a ontologia representa a reflexão sobre a realidade, que faz com que seja possível discernir objetos e compreender suas características e relações. Esta capacidade de discernimento e compreensão é a fundamentação da construção de uma ontologia computacional, onde se constrói uma estrutura que simula essa capacidade (Carvalheira, 2007).

Referências para o texto de Ontologias

CARVALHEIRA, Luiz Carlos da Cruz. **Método semi-automático de construção de ontologias parciais de domínio com base em textos**. Dissertação (Mestrado) – Escola Politécnica, Universidade de São Paulo, São Paulo, 2007. Disponível em <http://www.teses.usp.br/teses/disponiveis/3/3141/tde-10012008-094436/>. Acesso em: 11 abr 2008

GRUBER, Thomas R. **Toward principles for the design of ontologies used for knowledge sharing**. *International Journal of Human-Computer Studies*, v. 43, n. 5/6, p. 907-928, 1995. Disponível em: <http://www.uni-leipzig.de/~tbittner/courses/GEOID/ Grubner-Onto-design.pdf>. Acesso em: 01 jul 08.

WORDNET. **Ontology definition**. [S. l: s. n.], 2006. Disponível em: <http://wordnet.princeton.edu/>. Acesso em: 07 jul 08.

REIVINDICAÇÕES

1. Método para mapeamento de documentos textuais para bases de dados estruturadas utilizando ontologias **caracterizado por** compreender a seguinte sequência:

Fase 1: identificação de padrões e construção da ontologia com a definição da BD (TAV)

- Etapa 1: Identificação de padrões

a) Identificação de frases únicas: eliminar as frases repetidas formando um Conjunto de Frases Únicas (CFU); e

b) Redução de quantidade e de ambigüidade dos termos contidos no CFU:

i. Aplicação da técnica de *Stemming e/ou Lematização*;

ii. Definição da lista de *stopwords*;

iii. Construção do Arquivo de Padronização; e

iv. Geração de *n-gramas*;

- Etapa 2: Construção da estrutura para representação do conhecimento

a) Construção da ontologia; e

b) Definição da TAV;

Fase 2: mapeamento das informações contidas nos documentos digitais textuais para a BD (TAV)

- Etapa 1: Realização da padronização dos documentos originais; e

- Etapa 2: Processamento dos documentos padronizados por um Algoritmo de Mapeamento.

a) Divisão do documento em sentenças;

b) Processamento das sentenças; e

c) Preenchimento de atributos não-mapeados.

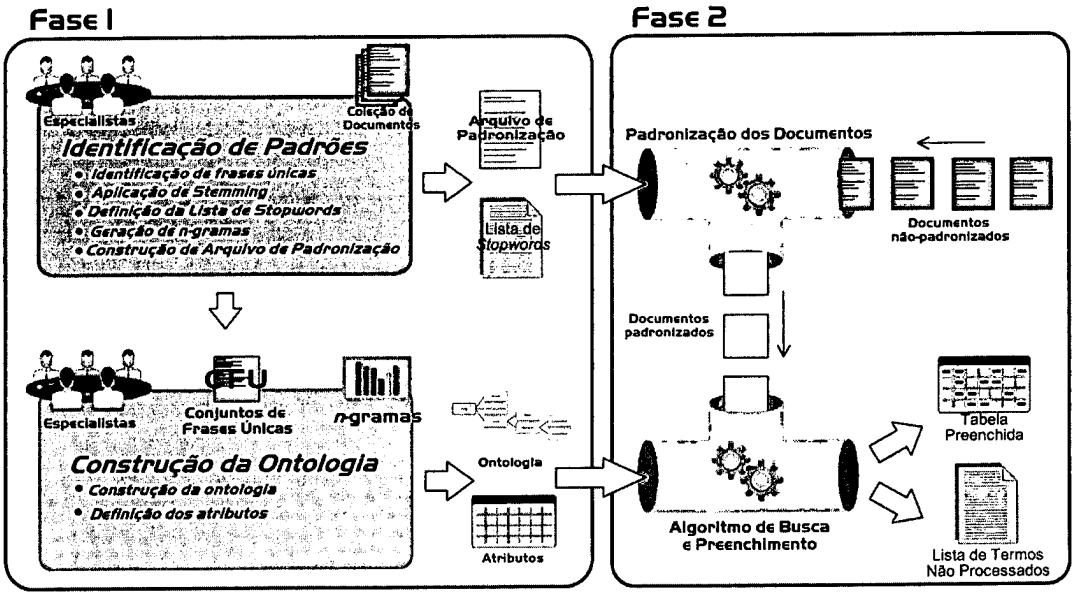

Figura 1

```
ESÔFAGO
- Mucosa de terço distal com presença de erosões, não confluentes
- Calibre e distensibilidade normais
- Motilidade normal
- TEG situada a aproximadamente 3,0 cm acima do pinçamento diafragmático
```

Figura 2

```
esofago
 esofago_inferior erosao_sim nao confluentes
 calibre normal
 distensibilidade normal
 motilidade normal
 teg gii
```

Figura 3

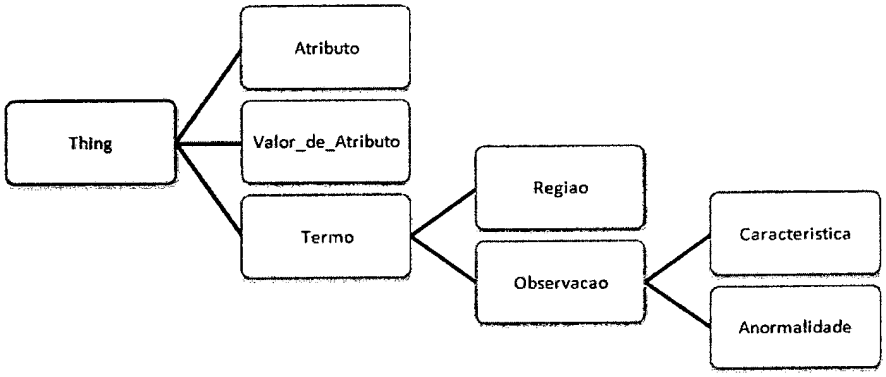
Thing
Atributo
Valor_de_Atributo
Termo
Regiao
Observacao
Caracteristica
Anormalidade

Figura 4