UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE CIÊNCIA DA INFORMAÇÃO

LUCÉLIA PINTO BRANQUINHO

MODELO PARA SUPORTE À DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (KDD): APLICAÇÃO EM ESTRATÉGIAS NO MERCADO DE MEDICINA DIAGNÓSTICA

Belo Horizonte
2015

LUCÉLIA PINTO BRANQUINHO

MODELO PARA SUPORTE À DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (KDD):
APLICAÇÃO EM ESTRATÉGIAS DE VENDA NO MERCADO DE MEDICINA DIAGNÓSTICA

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais para obtenção do grau de Mestre em Ciência da Informação.

Linha de Pesquisa: GIC - Gestão da Informação e do conhecimento

Orientador: Renata Maria Abrantes Baracho Porto

Co-orientador: Mauricio Barcellos Almeida

BELO HORIZONTE

2015

Branquinho, Lucélia Pinto.

B821m

Modelo para suporte à descoberta de conhecimento em base de dados (KDD) [manuscrito] : aplicação em estratégias no mercado de medicina diagnóstica / Lucélia Pinto Branquinho. – 2015.

120 f. : enc., il.

Orientadora: Renata Maria Abrantes Baracho.

Co-orientador: Maurício Barcellos Almeida.

Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

Referências: f. 90-97.

Apêndices: f. 98-120.

1. Ciência da informação – Teses. 2. Ontologias (Recuperação da Informação) – Teses. 3. Medicina – Diagnóstico – Teses. 4. Mineração de dados (Computação) – Teses. I. Título. II. Baracho, Renata Maria Abrantes. III. Almeida, Maurício Barcellos. IV. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.

CDU: 025.4:61

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG.

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

FOLHA DE APROVAÇÃO

"MODELO PARA SUPORTE À DESCOBERTA DE CONHECIMENTO EM BASE DE DADOS (KDD): APLICAÇÃO EM ESTRATÉGIAS NO MERCADO DE MEDICINA DIAGNÓSTICA"

Lucélia Pinto Branquinho

Dissertação submetida à Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, como parte dos requisitos à obtenção do título de "**Mestre em Ciência da Informação**", linha de pesquisa "**Gestão da Informação e do Conhecimento**".

Dissertação aprovada em: 31 de agosto de 2015.

Por:

Profa. Dra. Renata Maria Abrantes Baracho Porto - ECI/UFMG (Orientadora)

Prof. Dr. Renato Rocha Souza - FGV/RJ (por videoconferência)

Prof. Dr. Marcello Peixoto Bax - ECI/UFMG

Aprovada pelo Colegiado do PPGCI

Profa. Beatriz Valadares Cendón
Coordenadora

Versão final Aprovada por

Profa. Renata Maria Abrantes Baracho Porto
Orientadora

Av. Antônio Carlos, 6627 – Sala 2003 – Campus Pampulha – Cx. Postal 1606 – CEP: 30161-970 – Belo Horizonte – MG
Telefone: (31)3409-6103 – Fax: (31)3409-5207 – www.eci.ufmg.br/ppgci – e-mail: ppgci@eci.ufmg.br

UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-Graduação em Ciência da Informação

ATA DA DEFESA DE DISSERTAÇÃO DE **LUCÉLIA PINTO BRANQUINHO**, matrícula: 2013708917

Às 10:00 horas do dia 31 de agosto de 2015, reuniu-se na Escola de Ciência da Informação da UFMG a Comissão Examinadora aprovada *ad referendum* pelo Sub-coordenador do Programa de Pós-Graduação em Ciência da Informação em 06/08/2015, para julgar, em exame final, o trabalho intitulado ***Modelo para suporte à descoberta de conhecimento em base de dados (KDD): aplicação em estratégias no mercado de medicina diagnóstica***, requisito final para obtenção do Grau de MESTRE em CIÊNCIA DA INFORMAÇÃO, área de concentração: Produção, Organização e Utilização da Informação, Linha de Pesquisa: Gestão da Informação e do Conhecimento. Abrindo a sessão, a Presidente da Comissão, Profa. Dra. Renata Maria Abrantes Baracho Porto, após dar conhecimento aos presentes do teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata para apresentação de seu trabalho. Seguiu-se a argüição pelos examinadores com a respectiva defesa da candidata. Logo após, a Comissão se reuniu sem a presença da candidata e do público, para julgamento e expedição do resultado final. Foram atribuídas as seguintes indicações:

Profa. Dra. Renata Maria Abrantes Baracho Porto - Orientadora	APROVADA
Prof. Dr. Renato Rocha Souza (por videoconferência)	APROVADA
Prof. Dr. Marcello Peixoto Bax	APROVADA

Pelas indicações, a candidata foi considerada APROVADA.

O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a sessão, da qual foi lavrada a presente ATA que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 31 de agosto de 2015

Profa. Dra. Renata Maria Abrantes Baracho Porto
ECI/UFMG

Prof. Dr. Renato Rocha Souza
FGV/RJ

Prof. Dr. Marcello Peixoto Bax
ECI/UFMG

Profa. Beatriz Valadares Cendón
Coordenadora do Programa Pós-Graduação
em Ciência da Informação - ECI/ UFMG

Obs: Este documento não terá validade sem a assinatura e carimbo da Coordenadora

AGRADECIMENTOS

A Deus por iluminar e me dá força para alcançar esta vitória.

À minha família e amigos pela paciência e apoio incondicional. Sem eles nada disso seria possível.

Aos orientadores pelo apoio, incentivo e ensinamento.

Por fim, o meu profundo e sincero agradecimento a todos que contribuíram para a concretização deste sonho, estimulando-me intelectual e emocionalmente.

Mais um sonho se torna realidade...

“Thus, the task is not so much to see what no one yet has seen, but to think what nobody yet has thought about that which everybody sees.”

Arthur Schopenhauer

RESUMO

A grande quantidade de dados acumulados nos bancos de dados informatizados das organizações pode esconder conhecimentos valiosos e úteis para a tomada de decisão. A mineração de dados é uma das técnicas adotadas para identificar estes padrões. A área da saúde oferece inúmeras possibilidades de aplicações destas técnicas devido a complexidades dos processos e o grande volume de armazenamento de seus dados em uso pelos sistemas de informação. Em um domínio específico, um desafio dos processos de recuperação da informação é criar a relação semântica entre os termos de um vocabulário especializado, gerando um modelo de representação do conhecimento eficiente, através, por exemplo, de instrumentos como as ontologias.

O objetivo deste trabalho é descrever uma proposta de uso de ontologias biomédicas de doenças e testes laboratoriais no processo de *Knowledge Discovery in Database* (KDD) para tornar mais efetiva a recuperação da informação sobre o comportamento de prescrição de testes laboratórios, no caso deste experimento, relacionados às hepatites virais. Associa-se a técnica de organização de conhecimento e a de mineração de dados contribuindo para a recuperação da informação e, consequentemente, para extração do conhecimento.

O modelo desenvolvido instancia uma ontologia de testes complementares das hepatites virais para generalização dos termos do conjunto para hierarquia de nível mais alto na fase de pré-processamento, e posteriormente, classificar as regras de associação obtidas considerando a similaridade semântica entre o antecedente e consequente, ou seja, avaliar os testes complementares considerando a semelhança entre as doenças. As ontologias de domínio são utilizadas para representar conceitualmente os termos e introduz conhecimentos dos especialistas permitindo poda e classificação dos padrões filtrando itens mais interessantes.

Após a validação do modelo, foi criado um banco de dados com os pedidos de janeiro até março de 2015 que continham algum teste complementar diretamente relacionado ao diagnóstico de hepatites virais. Os registros a serem minerados foram generalizados por uma aplicação desenvolvida em Java com framework Jena e posteriormente, as regras de associação obtidas foram classificadas utilizando o modelo cálculo de similaridade de Tversky.

Os resultados obtidos nos experimentos mostraram que com a generalização dos atributos, consolidação de testes complementares relacionados, e, posteriormente, classificação das regras, avaliando as características relacionadas à doença, seus sintomas e o local onde predominante atacam, utilizando ontologia é possível reduzir o número de padrões gerados e, portanto, recuperar informações mais relevantes para a tomada de decisão. A análise dos resultados obtidos possibilita aos gestores direcionar de forma mais efetiva ações de divulgação dos ciclos de venda de marketing e abordagem de vendas.

Palavras-chave: KDD. Regras de associação. Ontologia. Medicina Diagnóstica.

ABSTRACT

The great deal of data stored in databases of healthcare organizations may mask valuable and useful possibilities in decision making. The health sector offers numerous possibilities for applications of these techniques due to the complexity of the processes and the large storage volume of your data in use by information systems. In a specific domain, a challenge of the information retrieval process is to create the semantic relation between the terms of a specialized vocabulary. A well-known alternative to identify hidden standards is the use of data-mining techniques.

In order to obtain more efficiency in data-mining, ontologies have been used to Such study aims at describing a proposal for the use of biomedical ontologies diseases and laboratory tests on the process with Knowledge Discovery in Database (KDD) to make more effective information retrieval on the prescribing behavior of laboratory tests which is, in this case, related to viral hepatitis. It is associated with such knowledge organizational technique and its data mining which contributes to the information retrieval and, consequently, to the knowledge gained.

The model developed shows ontology additional testing of viral hepatitis to generalize its attributes, in the pre-processing phase, and in the pos-processing phase and classify the association rules obtained considering the semantic similarity between the antecedent and consequent, assess additional testing considering what similarly relates to their disorders. The domain ontologies are used to introduce its theoretical terms and introduce experts' knowledge which allows its patterns control and assessment, considering the most interesting items.

After model validation, a database was created with the January to March of 2015 requests, which contained some additional testing directly related to the diagnosis of viral hepatitis. The records to be mined were widespread by an application developed in Java within Jena framework and subsequently the association rules obtained were classified using the Tversky similarity calculation model.

The results showed that with the generalization of attributes, the consolidation of related additional tests, and subsequently the classification of rules, evaluating the features related to the disease, its symptoms and where they prevailingly attack using ontology is possible to reduce the number of generated patterns and therefore recover more relevant information in decision making.

The analysis of the results allows managers to target more effectively their marketing sales cycles and sales approach.

Keywords: *KDD. Association rules. Ontology. Medical diagnosis.*

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS

AKD	*Actionable knowledge discovery and delivery*
ANS	Agência Nacional de Saúde Suplementar
BFO	*Basic Formal Ontology*
BI	*Business Intelligence*
CC	Ciência da Computação
CI	Ciência da Informação
DAML	*DARPA Agent Markup Language*
DAG	*Directed Acyclic Graph*
DM	*Data Mining*
DOID	Disease Ontology
DOLCE	Descriptive Ontology for Linguistics and Cognitive Engineering
ES	Engenharia de Software
GFO	*General Formal Ontology*
IA	Inteligência Artificial
IC	Contéudo da informação
IDO	*Infectious Disease Ontology*
IE	Inteligência Empresarial
FMA	*Foundation model of anatomy ontology*
HVO	*Ontologia Hepatite Viral*
KDD	*Knowledge Discovery in Database*
KOS	*Knowledge Organization Systems*
LHS	*Left Hand Side*
RHS	*Right Hand Side*
MD	Mineração de Dados

MeSH	*Medical Subject Headings*
MS	Ministério da Saúde
MOAL	Multi-ontology data mining at all Levels
OBO	*Open Biomedical Ontologies Foundry*
OGMS	*Ontology for general medical science*
OIL	*Ontology Inference Language*
OWL	*Web Ontology Language*
RDF	*Semantic Web Standards*
RI	Recuperação da Informação
SHOE	*Simple HTML Ontology Extensions*
SI	Sistemas de informação
SIC	Sistema de Inteligência Competitiva
SNOMED CT	*Systematized Nomenclature of Medicine--Clinical Terms*
SOC	Sistema de Organização do Conhecimento
SRI	Sistema de Recuperação da Informação
SUS	Sistema Único de Saúde
TOVE	*Toronto Virtual Enterprise*
UFO	*Unified Foundational Ontology*
UMLS	*Unified Medical Lingual System*
XML	*Extensible Markup Language*
XOL – XML	*Based Ontology Exchange Language*
W3C	*World Wide Web Consortium*

SUMÁRIO

1 INTRODUÇÃO

A quantidade de informações armazenadas em bancos de dados das organizações está ultrapassando a habilidade técnica e a capacidade humana na sua interpretação. Tais constatações mostram a necessidade de ferramentas que sejam capazes de analisar automaticamente as bases de dados para obter conhecimento (DALFOVO; AMORIM, 2000).

As organizações precisam obter informações úteis aos seus objetivos para criar estratégias adequadas e promover a inovação agregando valor à tomada de decisão. A inteligência competitiva é característica marcante nas empresas atuais e seu sucesso é decorrente da gestão da informação e do conhecimento produzido, interno ou externo. O grande desafio é estruturar a recuperação da informação para obter conhecimento relevante.

Quoniam (2000) mostra, na Figura 1, o posicionamento lógico de diferentes fases da tomada de decisão com seu valor potencial para as dimensões tática e estratégica. O valor estratégico da informação aumenta quando os dados estão altamente resumidos.

FIGURA 1 - Mineração de dados e Inteligência de negócios

Fonte: HAN; KAMBER, 2006, p. 12.

Uma das técnicas para extração do conhecimento geralmente referenciada na literatura é o *Knowledge Discovery in Database* (KDD)

No entanto, a comunidade de mineração de dados se confronta com o desafio de explorar o grande volume de dados utilizando recurso de conhecimento de domínio dos dados semanticamente anotados para obter maior precisão e revocação na descoberta de conhecimento. Vavpetic (2012) referência o termo ‘mineração de dados semânticos’ para designar a nova mineração de dados que muda com a introdução destas abordagens.

A aplicação de métodos e tecnologias para mineração de dados semântica permite que informações estratégicas não explícitas dos domínios representados sejam descobertas e transformadas em conhecimento relevante para compreensão de problemas complexos e no processo de tomada de decisão. Ao integrar métodos estatísticos com abordagem semântica é possível identificar com sucesso padrões não triviais, semanticamente corretos, mais relevantes e específicos (FERRAZ, 2008).

Para o mercado de medicina diagnóstica é importante entender o comportamento de prescrição dos médicos no diagnóstico das doenças para antecipar tendências e assim realizar ações de marketing e vendas direcionadas ao mercado. Portanto, entender o domínio dos testes laboratoriais complementares ao diagnóstico possibilita aprimorar os padrões extraídos no processo de mineração de dados.

Considerando a necessidade do mercado de medicina diagnóstica, este estudo tem como objetivo aprimorar o processo de mineração de dados através da introdução do conhecimento do domínio na fase de pré-processamento e pós-processamento. A extração de conhecimento com adoção de ontologias biomédicas permite incorporar à mineração de dados medidas subjetivas para generalização e classificação com base na hierarquia e relações que estes termos compartilham trazendo maior efetividade e relevância na classificação dos padrões de regras de associação.

Como delimitação para este experimento foi considerado o universo dos exames laboratoriais de análises clínicas para diagnóstico das hepatites virais humanas sendo considerado como referência para codificação o padrão *Logical*

Observation Identifiers Names and Codes[1][2] (LOINC), definido como padrão de interoperabilidade de sistema de informação pela portaria nº 2073 de 31 de agosto de 2011[3] do Ministério da Saúde Brasileiro.

Considerando os protocolos do Ministério da saúde sobre hepatites virais, a relação de testes laboratoriais do LOINC e a pesquisa em um laboratório de medicina diagnóstica foi possível, com o reuso das ontologias *Ontology for General Medical Science*[4] (OGMS), *Disease Ontology*[5] (DOID) e *Foundational Model of Anatomy*[6] (FMA), mapear o conhecimento sobre as hepatites virais, conforme apresentado no apêndice C. Este mapeamento possibilitou a generalização dos termos para hierarquia de nível mais alto e, consequente, redução no número de atributos a serem minerados, além da classificação dos padrões obtidos através do cálculo de similaridade entre o antecedente e consequente. A generalização dos testes complementares associados ao diagnóstico de hepatites virais e, posterior classificação dos padrões de comportamento de associação entre eles possibilitaram recuperar resultados mais relevantes como os apresentados nas figuras de 21 à 24.

Este trabalho tem como objetivo geral: Identificar como o uso de ontologias biomédicas pode aprimorar o processo de *Knowledge Discovery in Database* (KDD) por regras de associação em uma empresa de medicina diagnóstica.

No transcorrer do trabalho, pretende-se atingir os seguintes objetivos específicos:

- Ilustrar como as ontologias biomédicas podem apoiar o processo de KDD por regras de associação em empresas do mercado de medicina diagnóstica;
- Apresentar aos especialistas do negócio a importância do uso de sistemas de organização do conhecimento para maior eficiência da extração de conhecimento para tomada de decisão;

---

[1] Disponível em:<https://loinc.org/>. Acesso em: 1 mar. 2014.

[2] Disponível em:<http://www.ihtsdo.org/snomed-ct>. Acesso em: 1 mar. 2014.

[3] Disponível em:<http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html>. Acesso em: 1 mar. 2014.

[4] Disponível em:<http://www.obofoundry.org/cgi-bin/detail.cgi?id=OGMS>. Acesso em: 1 mar. 2014.

[5] Disponível em:<http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology>. Acesso em: 1 mar. 2014.

[6] Disponível em:<http://sig.biostr.washington.edu/projects/fm/>. Acesso em: 1 fev.2015.

- Avaliar se os padrões obtidos com o processo de KDD contribuem na validação e expansão das ontologias biomédicas.

As próximas seções serão organizadas da seguinte maneira: no Capítulo 2 faço uma breve revisão de literatura sobre: descoberta de conhecimento em base dados (KDD), ontologias, ontologias e KDD por regras de associação e similaridade semântica com o uso de ontologias e relata alguns trabalhos que utilizaram ontologias para aprimorar o processo de recuperação por regras de associação. O Capitulo 3 é dedicado a detalhar a metodologia para desenvolvimento do experimento e descreve brevemente a organização no qual foi validado. Já o Capitulo 4 detalha os resultados obtidos. No Capitulo 5, relato as considerações advindas deste experimento e menciono os trabalhos futuros.

2 MARCO TEÓRICO

Este capítulo tem como objetivo apresentar um panorama geral dos assuntos: ontologia, descoberta de conhecimento em base de dados (KDD), linguagens e ferramentas de mineração de dados e ontologia, modelos de integração de ontologias e KDD utilizando a técnica de mineração de regras de associação e similaridade semântica entre os termos. A metodologia utilizada para a realização da revisão foi através de levantamento bibliográfico. Foram utilizados artigos e outras referências apontadas nas disciplinas do mestrado e feitas pesquisas bibliográficas nas bases: Scielo, Portal Capes, CitseerX, BRAPCI e Pubmed, no período de setembro de 2013 a dezembro de 2014, utilizando descritores como: ontologias; similaridade semântica, descoberta de conhecimento em base de dados; ontologias e mineração de dados; ontologias, mineração de dados e similaridade semântica. Foram selecionados como direcionador do estudo os artigos, teses e dissertações listados no quadro 1.

Quadro 1 – Marco teórico

Assuntos	Autores
Descoberta de conhecimento de base de dados (KDD)	Fayyad (1996); Chapman (2000); Maimon e Rokach (2005); Ham e Kamber (2006); Gonçalves (2005); Souza e Carvalho (2007); Ferraz (2008);
Regras de associação	Agrawal (1993); Hasher (2007); Ferraz (2008); Gonçalves (2011); Ribeiro (2010); Camilo (2010); Coelho (2012)
Ontologias	Gruber (1993); Guarino e Giaretta (1995); Vickery (1997); Guarino (1998); Soergel (1999); Wand e Weber (1999); Noy e MCGuinness (2001), Almeida e Bax (2003); Smith (2004); Freitas (2005); Lima (2005); W3C (2009); Almeida (2010); Coelho (2010); Almeida (2013)
Similaridade semântica em ontologias	Tversky (1977); Resnik (1999); Baader (2003); Almeida (2011); Breitman (2005); Pesquisa (2009); Almeida, Souza e Fonseca (2011); Couto (2011); Gelaim (2013); Harispe (2013); Gan (2013)
Ontologias e KDD	Nigro et al (2007); Ferraz (2008); Marinica (2009); Camilo (2010); Ribeiro (2010); Cao (2010); Camilo (2010); Yokome (2010); Manda (2010, 2013); Coelho (2011, 2012); Vavpetic (2012); Hamani (2013, 2014);
Trabalhos relacionados	Ferraz (2008); Vivacqua (2008); Marinica (2010); Ribeiro (2011); Coelho (2012); Vavpetic (2012); Manda (2013); Hamani (2014); Harispe (2013); Gelaim (2013)

2.1 Descoberta de conhecimento de base de dados (KDD)

A extração de conhecimento, geralmente referenciada na literatura como *Knowledge Discovery in Database* (KDD) é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como banco de dados, inteligência artificial e estatística.

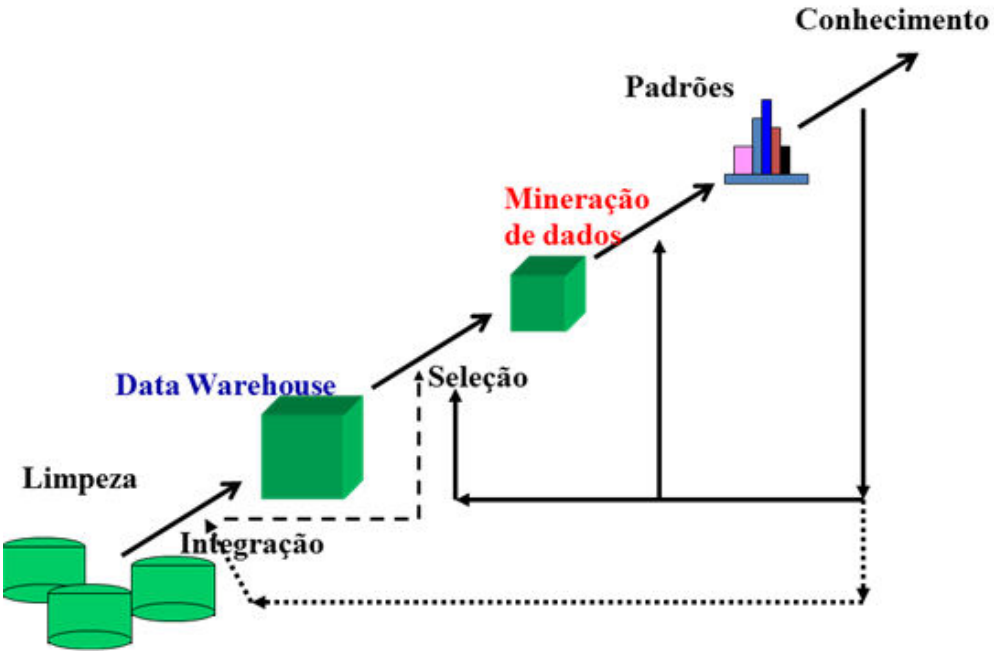
O uso de ferramentas para a descoberta de conhecimento em banco de dados (KDD) vem integrar a tecnologia e a aprendizagem organizacional em busca da gestão estratégica do conhecimento.

2.1.1 Definições e características

Segundo Fayyad *et al.* (1996), o processo de Extração de Conhecimento em Base de Dados tem como objetivo identificar padrões válidos, novos, potencialmente úteis e compreensíveis através da descoberta de associações e padrões entre os dados como, agrupamentos, classificações e outras avaliações e análises possibilitadas com o uso de algoritmos executados no decorrer do processo.

A Figura 2 representa KDD sobre a perspectiva de Han e Kamber (2006). Para um melhor entendimento o processo de KDD pode ser agrupado em três fases: pré-processamento, mineração de dados (MD) e pós-processamento. O pré-processamento compreende a captação, organização e tratamento dos dados; já a MD, os algoritmos e as técnicas para busca dos padrões; o pós-processamento abrange o resultado obtido na MD e a sua interpretação.

FIGURA 2 – Processo de descoberta de conhecimento (KDD)



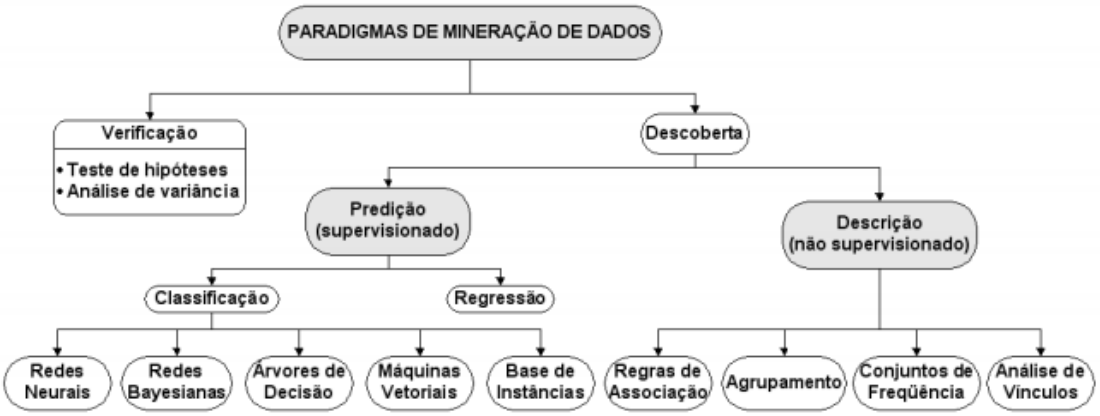
Fonte: HAN;KAMBER, 2006, p.12.

O processo de KDD não se resume apenas à MD, mas o uso destas técnicas é elementar para geração do conhecimento a partir de bases de dados.

Para Fayyad *et al.* (1996) a mineração de dados é uma parte do KDD responsável pela análise dos dados para obtenção de padrões através da aplicação de algoritmos.

Conforme Figura 3, a MD possui várias técnicas que variam de acordo com o objetivo a ser alcançado. A descoberta de padrões pode ser pré-definida, ou seja, supervisionada com características prescritivas e preditivas ou não supervisionada, onde não há conhecimento prévio, com características descritivas.

FIGURA 3 - Taxonomia de paradigmas de MD



Fonte: MAIMON; ROKACH, 2005, p.7.

Métodos atuais de MD diferem quanto ao tipo de informação extraído, indo desde regras de associação até padrões ou classificações. Em todas as estratégias, o objetivo maior é o de poder generalizar o conhecimento adquirido para novas ocorrências do fenômeno ou para outros contextos ou situações parecidas com a utilizada na construção do modelo computacional. Segundo Ribeiro (2010), a maioria das técnicas usadas no processo de MD necessita de um conjunto reduzido de atributos, pois a diminuição na quantidade de atributos possibilita maior rapidez e eficiência, permitindo uma classificação mais criteriosa e resultados mais fáceis de serem interpretados. A etapa de seleção de atributos tem como objetivo a identificação de quais informações, dentre as existentes na base de dados, devem ser consideradas para os próximos passos do processo de KDD.

Na etapa de pós-processamento, dentre as medidas utilizadas para se avaliar os padrões resultantes da MD há as medidas de interesse objetivas (*data-drive*) e subjetivas (*user-driven*) que buscam avaliar a novidade do padrão descoberto (SOUZA; CARVALHO, 2007; GONÇALVES, 2005).

Há duas principais medidas subjetivas que devem ser consideradas (HAMANI, 2014):

- Não previsibilidade: regras interessantes são desconhecidas pelo usuário ou contradizem o conhecimento prévio/expectativas;
- Acionáveis: regras que possibilitam uma ação imediata para obter alguma vantagem.

Para o sucesso na implantação do processo de KDD há necessidade do entendimento do problema, conhecimento do domínio da aplicação, que só será possível com a participação do usuário final. Dessa forma, a interação entre os usuários do processo (usuários finais, especialistas do domínio e analistas do processo) é imprescindível para o sucesso do projeto. Em função desta necessidade, um novo paradigma para a descoberta de conhecimento tem sido proposto para aplicação de ontologias com base em mineração guiada pelo conhecimento de domínio (*domain-driven data mining*, D3M) e pela praticabilidade de uso do conhecimento que é descoberto e entregue (*Actionable Knowledge Discovery and Delivery*, AKD) (MARINICA, 2010; RIBEIRO, 2010; CAO, 2010; CAMILO, 2011; YOKOME, 2011; COELHO, 2012).

Para Luz (2003) a MD é oposta à recuperação de informação, pois busca revelar alguns padrões ocultos, explorar seus dados e descobrir algumas características que não são perceptíveis em uma análise mais superficial.

Em geral, o conhecimento descoberto através de processos de KDD é expresso na forma de regras e padrões. Na busca de padrões são escolhidos técnicas e algoritmos. A escolha da técnica de MD apropriada depende do tipo de tarefa de KDD a ser realizada. Dentre os diferentes tipos de técnicas que podem utilizadas para extração em bases de dados encontram-se as regras de associação, classificação entre outros. A descoberta de regra de associação que procura itens que ocorrem frequentemente em transações do banco de dados. O exemplo clássico é o de “cestas de compras” na área de marketing com a intenção de estimular a compra de itens que são comprados juntos.

2.1.2 Mineração de dados por regras de associação

A mineração de dados por regras de associação é uma das técnicas mais utilizada sendo a tarefa de associação realizada por meio de algoritmos que geram regras que caracterizam o quanto a presença de um conjunto de itens, nos registros de uma base de dados, implica na presença de outro conjunto distinto de itens, nos mesmos registros (FERRAZ, 2008).

A técnica de mineração por regras de associação busca correlação entre conjuntos de itens frequentes em uma série de dados ou transações. A partir de

conjuntos de elementos que aparecem juntos com pelo menos alguma frequência (suporte), as regras de associação representam a condição "SE antecedente ENTÃO consequente" com garantia probabilística (confiança), que sempre que o antecedente ocorrer o consequente também estará presente (FERRAZ, 2008). Uma regra de associação é uma implicação da forma A=> B, para A e B conjuntos disjuntos de itens de dados.

As medidas objetivas, como o suporte e confiança, são usadas para medir a importância de uma regra de associação.

Seja I = {i1, i2, i3, ..., in-1, in}. um conjunto de dados. Transação é um conjunto de itens T= {t1, t 2, t 3,..., t n-1, t n}, T ⊂ I. D é um conjunto de transações ou dados relevantes para a tarefa. Uma regra de associação é uma implicação de forma P → Q onde P ⊂ I, Q ⊂ I e P∩Q = ∅ (Ferraz, 2008).

O suporte de uma regra P → Q é definido como o percentual de transações da base de dados em que o antecedente e consequente da regra aparecem na mesma transação:

$$\text{Suporte } (P \rightarrow Q) = \frac{\text{total de transações com ocorrências } (P \cup Q)}{\text{total de transações}}$$

A confiança de uma regra P → Q é definida como o percentual de transações, dentre as que possuem o antecedente da regra, em que antecedente e consequente aparecem conjuntamente na mesma transação.

$$\text{Confiança } (P \rightarrow Q) = \frac{\text{total de transações com ocorrências } (P \cup Q)}{\text{total de transações com ocorrências P}}$$

Uma regra P → Q [s,c] é chamada de regra forte, se dados valores para os parâmetros suporte mínimo especificado(s) e confiança mínima especificada(c), então (FERRAZ, 2008):

sD(P → Q) = s, s ≥ suporte mínimo especificado

cD(P → Q) = c, c ≥ confiança mínima especificada

A tarefa de mineração por regras de associação pode ser dividida em dois passos (FERRAZ, 2008):

1) Extração de todos os itens frequentes, ou seja, todas as combinações de itens com suporte maior que mínimo definido pelo usuário;

2) A partir dos itens frequentes, extrair todas as regras com confiança maior que a definida pelo usuário, ou seja, para todos os subconjuntos não vazios

de cada item frequente x calcular a confiança c da regra s → (x-s) da forma c = sup({x,(x-s)}) / sup ({s}) = sup({x}) / sup ({s}). Se c ≥ maior que confiança mínima então aceitar a regra s → (x-s).

O suporte mínimo garante a relevância estatística da amostra, evitando a ocorrência de pequena frequência. Já a confiança mínima, garante que o resultado obtido não é ocasional havendo coesão entre as regras obtidas. Uma regra é considerada forte quando tem regularidade e confiança alta para uma grande quantidade de instâncias, enquanto uma regra fraca apresenta para uma pequena quantidade de instâncias (FERRAZ, 2008)

Para mensurar a dependência entre os itens é utilizada a métrica (*Lift*) (GONÇALVES, 2011). É uma métrica utilizada para avaliar dependências entre antecedente e consequente, quanto maior o valor do *lift*, mais interessante a regra, pois A aumentou (*"lifted")* B numa maior taxa.

Seja D uma base de dados de transações. Seja A => B uma regra de associação obtida a partir de D. Dada uma regra de associação A => B, esta medida indica o quanto mais frequente torna-se B quando A ocorre. O valor do *lift* para A => B é computado por:

Lift(A => B) = Conf(A =>B)/Sup(B)

Se Lift(A => B) = 1, então A e B são independentes.

Se Lift(A => B) > 1, então A e B são positivamente dependentes.

Se Lift(A => B) < 1, A e B são negativamente dependentes.

Um problema clássico da mineração de dados por regras de associação é a escolha da precisão dos resultados, pois com parâmetros pequenos para os valores de suporte e confiança o usuário pode ter um grande número de padrões e associações muito parecidos entre si. Com parâmetros com valores maiores, alguns padrões e associações interessantes podem desconsiderados.

Uma opção para resolver o problema da relevância dos dados obtidos é reduzir os dados a serem minerados através de pesquisa de representações dos conjuntos frequentes e regras "fortes" mais coesas. A redução do conjunto das regras pode ser feita pelo mecanismo de poda. A questão é: Quais as regras que devem ser podadas?

Segundo Ferraz (2008), há duas abordagens para poda das regras:

1) Medidas objetivas com uso de técnicas estatística;

2) Medidas subjetivas com dependência da ação do usuário, pois depende do conhecimento do domínio. Os resultados obtidos são significativamente melhores, mas a automatização do processo se torna inviável.

Para isso, é importante resolver o problema da dependência do conhecimento do domínio do especialista e tornar este processo mais automatizado. Portanto, os conceitos devem ser mapeados e representados de forma que possibilitem o processamento por computadores reduzindo a dependência da presença do especialista do domínio.

Um dos grandes problemas relacionados ao mapeamento e uso do conhecimento é a dificuldade para aquisição, pois muita investigação deve ser feita para elucidar um domínio, muitas fontes precisam ser pesquisadas e os conhecimentos podem não ser confiáveis devido à falta de formalismo. A Ciência da Informação (CI) contribui com a recuperação da informação ao introduzir técnicas para representação do conhecimento como as ontologias.

As ontologias são utilizadas para uma melhor compreensão sobre a natureza dos objetos do domínio a serem estudados permitindo a criação e formalização do conhecimento. Com a adoção de uma série de restrições nos relacionamentos entre conceitos que possibilita a identificação de generalizações / especializações e poda das regras de associação tornando mais eficiente o KDD. Com esta simplificação na definição e no uso de restrições, reduzem-se as diferenças entre regras descobertas e expectativas dos usuários.

2.2 Ontologias

As ontologias tem sido objeto de estudo em diversos campos de pesquisa, como por exemplo, na filosofia, ciência da computação e ciência da informação com abordagens direcionadas aos domínios da medicina, biologia, engenharia, geografia e direito, portanto, é um tema interdisciplinar (ALMEIDA, 2013).

Nos últimos anos, a pesquisa em ontologia tem recebido destaque em função das possibilidades que ela oferece para a representação e organização da informação. No campo da Ciência da Informação, trabalhos voltados para a pesquisa em ontologia com o foco para melhoria da representação formal de um domínio do conhecimento tem se destacado.

De acordo com Smith (2004), na década de 60, o termo Ontologia, começou a ser usado no campo da Ciência da Computação (CC), por Mealy[7] que o vinculou em sua pesquisa de Representação do Conhecimento, na subárea da Inteligência Artificial. Na década de 80, Wand e Weber (1999) pesquisaram a aplicação de ontologias no desenvolvimento de Sistemas de Informação (SI) iniciando as primeiras do seu uso na área de Engenharia de Software (ES). Na década de 90, segundo Vickery (1997), as publicações de pesquisadores da CI demonstravam um crescente interesse pela disciplina Ontologia.

Em CC há inúmeras publicações do uso de ontologias na representação de artefatos da engenharia de software, modelagem de sistema de banco de dados, e sistema de representação de conhecimento e inteligência artificial (GRUBER,1993).

Seguindo Guarino e Giaretta (1995) nos referimos ao termo ontologia, com um artigo indeterminado e uma minúscula inicial, como um objeto particular e Ontologia, sem artigo e maiúscula inicial, para referenciar a parte da filosofia que estuda a organização e a natureza do mundo.

Neste sentindo, as seções a seguir têm como principal objetivo apresentar a definição dada ao termo, suas principais características, tipos, metodologia e ferramentas para sua construção.

2.2.1 Definições e características

Presente inicialmente na filosofia em trabalhos de Aristóteles, a palavra Ontologia tem origem grega onde Onto quer dizer “ser” e logia “estudo”, assim “estudo do que é” (ALMEIDA; BAX, 2003). A Ontologia é considerada um ramo da metafísica[8] que estuda as categorias de entidades que existem e como cada uma delas está relacionada. O objetivo é dar sentido ao mundo, de seus objetos e às relações entre objetos (ALMEIDA, 2013).

Segundo *World Wide Web Consortium* (W3C) (2009), considerando o uso em CC, as ontologias são um conjunto de proposições legíveis por máquina que designam uma taxonomia de classes e subclasses e os relacionamentos entre elas.

---

[7] MEALY, G. H. Another look at data. In: AFIPS CONFERENCE, 31th, 1967. **Proceedings...** Washington, 1967. p. 525-534.

[8] Metafísica: ramo da filosófica que estuda a realidade.

De acordo com Gruber (1996) onde “uma ontologia é uma especificação explícita de uma conceitualização”. Segundo o autor, uma conceitualização é um grupo de relações extensionais descrevendo um problema em particular, ou é uma relação intencional que descreve o domínio geral do problema. Almeida e Bax (2003, p.9) classificam esta definição como simples e completa. Para eles, “formal” remete a legibilidade para computadores; “especificação explícita” remete a conceitos, propriedades, relações, funções restrições, axiomas, claramente definidos; “compartilhado” diz respeito ao conhecimento consensual; e “conceitualização” refere-se a um modelo abstrato de algum acontecimento do mundo real.

Muitas pesquisas têm sido produzidas acerca da contribuição da Ontologia tanto na CI quanto nas suas áreas multidisciplinares. Em relação à construção de ontologias, a fase de conceitualização, requer muita atenção, pois é quando o responsável pela organização do vocabulário representacional executa abstrações a fim de representar parte da realidade, importantes para seus objetivos. Neste ponto, a CI tem grande contribuição para a construção de ontologias a partir de suas teorias de análise conceitual e das relações semânticas.

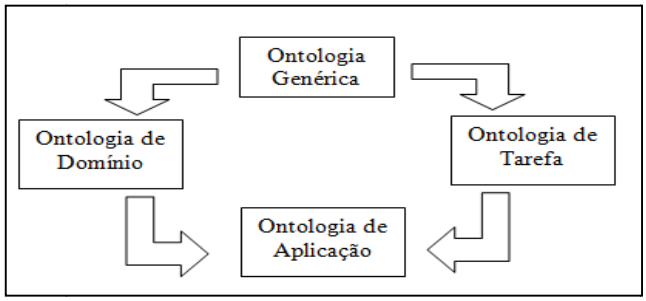
Almeida *et al.* (2010) cita vários autores que escrevem sobre as principais ontologias de referência ou de fundamentação. Podemos citar algumas como: a *Basic Formal Ontology* (BFO); a *Descriptive Ontology for Linguisitics and Cognitive Engineering* (DOLCE); a *General Formal Ontology* (GFO); e a *Unified Foundational Ontology* (UFO).

Vickery (1997) aponta que para construir ontologias os conceitos podem ser coletados através de consulta à literatura do domínio ou ao especialista do tema pesquisado. Neste trabalho, foram considerados ambos sendo validados com os especialistas (médico, biomédicos, farmacêuticos) os conceitos obtidos, conforme questionário constante do Apêndice A.

Guarino (1998) propõe que as ontologias em geral podem ser classificadas em ontologias genéricas, ontologias de domínio, ontologias de tarefa e ontologias de aplicação. As ontologias genéricas descrevem conceitos gerais, tais como tempo, espaço, matéria, entre outros, portanto, não estão relacionadas a um domínio específico. Já, as ontologias de domínio descrevem domínios específicos com intuito de evitar contradições conceitos. As ontologias de tarefa se propõem a descrever uma funcionalidade de um domínio através de sua atividade ou tarefa. Em

contrapartida, as ontologias de aplicação descrevem conceitos dependentes do domínio e da tarefa. A Figura 4 reforça os conceitos dados por Guarino (1998).

FIGURA 4 - Relação entre os tipos de ontologias



Fonte: GUARINO, 1998, p.9.

De acordo com Gruber (1993) os componentes básicos de uma ontologia são: i) classes – entidades organizadas taxonomicamente que representam o conjunto dos indivíduos, representando concretamente os conceitos; ii) relações – o tipo de interação entre os conceitos de um domínio; iii) Axiomas - usados para modelar sentenças sempre verdadeiras; iv) instâncias – utilizadas para representar elementos específicos, ou seja, os próprios dados.

Os axiomas são utilizados para modelar regras assumidas como verdadeiras no domínio em questão, de modo que seja possível associar o relacionamento entre os indivíduos, além de fornecer características descritivas e lógicas para os conceitos. Para Uschold e Grüninger (1996), os axiomas são especificados para definir a semântica e significado dos termos (classes e propriedades).

Representações de conhecimento, como as ontologias, são utilizadas para mapear o conhecimento do mundo possibilitando o uso e manipulação por máquinas permitindo, por exemplo, inferir novos conhecimentos sobre o domínio pesquisado.

2.2.2 Construção das ontologias

Há diferentes linguagens de representação e metodologias para construção de ontologias que proporcionam diferentes facilidades e recursos. Uma listagem mais completa de linguagens de representação de ontologias pode ser encontrada (ALMEIDA; BAX, 2003).

Neste trabalho, optou-se por trabalhar com a metodologia *Ontology Development* 101 (NOY; MCGUINNESS, 2001) por ser um guia de passos iterativos, que pode ser livremente executado sem muitas restrições e exigências, tornando-se uma metodologia simples para uso nesta pesquisa que não tem como foco a construção de uma ontologia de referência.

Esta metodologia aborda um processo iterativo com sete passos para construção de uma ontologia: (a) determinar o domínio e escopo da ontologia, (b) considerar o reuso de ontologias existentes, (c) enumerar termos importantes na ontologia, (d) definir as classes e a hierarquia das classes, (e) definir as propriedades das classes, (f) definir as restrições e por fim, (g) criar instâncias.

A modelagem ontológica através da *Web Ontology Language* (OWL) é a mais utilizada atualmente. A OWL é uma linguagem de definição de ontologias utilizada para definir termos e seus relacionamentos, recomendada pela W3C, e projetada para uso por aplicações que precisem processar o conteúdo de informação, em vez de apenas apresentá-la. A sintaxe OWL é baseada em *Extensible Markup Language* (XML) e *Semantic Web Standards* (RDF), compatível com RDFs (W3C, 2009). OWL surge no contexto da Web Semântica para permitir a representação de termos em vocabulários e seus inter-relacionamentos em uma ontologia.

O OWL formaliza um domínio, definindo classes e propriedades destas classes, indivíduos e afirmações sobre eles e, usando-se a semântica formal OWL, especificar como derivar consequências lógicas, isto é, fatos que não estão presentes na ontologia, mas são vinculados pela semântica (W3C, 2009).

A OWL permite que sejam impostas restrições sobre propriedades. Uma restrição é um tipo especial de descrição de classe, isto é, descreve uma classe anônima de indivíduos que satisfazem as restrições. As restrições podem ser de valores (*allValuesFrom*, *someValuesFrom* e *hasValue*) ou de cardinalidade (*maxCardinality*, *minCardinality* e *Cardinality*) (LIMA, 2005).

Duas classes podem ser equivalentes, ou seja, podem possuir exatamente a mesma extensão de classe. Esta construção pode ser usada para criar classes sinônimas e para ligar duas classes na mesma ontologia ou em ontologias diferentes. A construção OWL: *equivalentClass* não implica em igualdade de classes, ou seja, duas classes são iguais se elas possuem o mesmo significado

intencional. A OWL permite combinações booleanas arbitrárias para manipulações das extensões de classes, chamados de conjunto de operadores. Este conjunto de operadores podem ser visualizados como uma representação dos operadores *AND*, *OR* e *NOT*, usados na Lógica Descritiva, para relacionar as classes (LIMA, 2005).

A OWL possui três sub-linguagens (espécies):

- OWL-*Lite* - restrições e uma modelagem de hierarquia de classes simples;
- OWL-DL - lógica descritiva verifica inconsistências na linguagem ontológica, classificação automática;
- OWL-*Full* - maior dinamismo da linguagem, sem inferências.

Cada uma destas sub-linguagens é uma extensão de sua predecessora, ou seja, cada ontologia válida em OWL *Lite* é uma ontologia válida em OWL DL, esta por sua vez é uma ontologia válida em OWL *Full* (W3C, 2009).

O OWL DL é linguagem de representação de conhecimento que além de armazenar conceitos e afirmações pode realizar checagem de consistência e podem inferir conhecimento obtido através de inferências lógicas. Isto só é possível devido à utilização da lógica descritiva o qual possibilita automaticamente computar hierarquias de classes verificando suas inconsistências na linguagem ontológica.

Baader e Nutt (2003) definem a lógica descritiva como um formalismo matemático que representa o conhecimento de um domínio de aplicação (o “mundo”) primeiramente definindo os conceitos relevantes desse domínio (sua terminologia) e depois usando esses conceitos para especificar propriedades de objetos e indivíduos que ocorrem nesse domínio (a descrição do mundo). Esse formalismo é que garante que ontologias em OWL DL sejam computáveis.

Uma ontologia OWL DL é equivalente a uma base de conhecimento de Lógica Descritiva onde os conceitos são chamados de classes e papéis são chamados de propriedades (GELAIM, 2013).

A ferramenta capaz de gerar as inferências[9] lógicas a partir dos axiomas existentes e checar a consistência em OWL é chamada de *reasoner* ou motor de inferência ou racionador.

[9] Inferência é a derivação de novas sentenças a partir de sentenças antigas (RUSSEL, 2006)

Uma das vantagens do uso da linguagem OWL DL na construção de ontologias é que os racionadores, por exemplo, HermiT[10], Pellet[11], FaCT++[12] proporcionam um suporte genérico, aplicável a todos os domínios de aplicação.

Para ontologias construídas em OWL pode-se utilizar racionadores como mecanismos de inferência para busca das regras da base de conhecimento a serem avaliadas, direcionando o processo de inferência, inserindo novas informações, classificando as instâncias ou verificando regras e consistência. As definições e seus relacionamentos são formalizados em axiomas lógicos, o que permite agregar inferência lógica à ontologia e desenvolver mecanismos de classificação (COELHO, 2010).

O conhecimento em lógica descritiva é dividido em duas entidades: TBox contendo o conhecimento taxonômico, classes e suas hierarquias, e o ABox constituído de asserções sobre indivíduos (relações entre os indivíduos e conceitos) feitas utilizando os conceitos e papéis definidos pelo TBox (BADDER, 2003). O raciocínio em sistema de lógica descritiva pode ser sobre conceitos TBox e ABox o qual são equivalentes semanticamente a um conjunto de axiomas.

Sobre esse conhecimento é possível explicitar conhecimento implícito, ou seja, raciocinar. Por exemplo, é possível avaliar a similaridade entre dois conceitos definidos no TBox considerando tanto a estrutura da representação quanto as características dos conceitos comparados.

Investigações recentes têm enfatizado o uso de medidas de semelhança semântica como um mecanismo valioso para recuperação de informação e descoberta de conhecimento, pois permite avaliar a semelhança dos conceitos definidos em ontologias considerando sua estrutura taxonômica (HARISPE, 2013).

2.2.3 Similaridade semântica

A similaridade atua como uma forma de organizar, classificar, construir conceitos e generalizações (TVERSKY, 1977), podendo assim, ser considerada uma forma de raciocínio. Segundo (GELIAM, 2013) há dois aspectos importantes para

[10] Disponível em:< http://www.hermit-reasoner.com/java.html>. Acesso em: 1.fev.2015.

[11] Disponível em:< https://github.com/complexible/pellet>. Acesso em: 1.fev.2015.

[12] Disponível em:< http://owl.man.ac.uk/factplusplus/>. Acesso em: 1.fev.2015.

medição de similaridade: um qualitativo, para avaliar em que dois objetos são similares, e outro quantitativo, quanto dois objetos são similares.

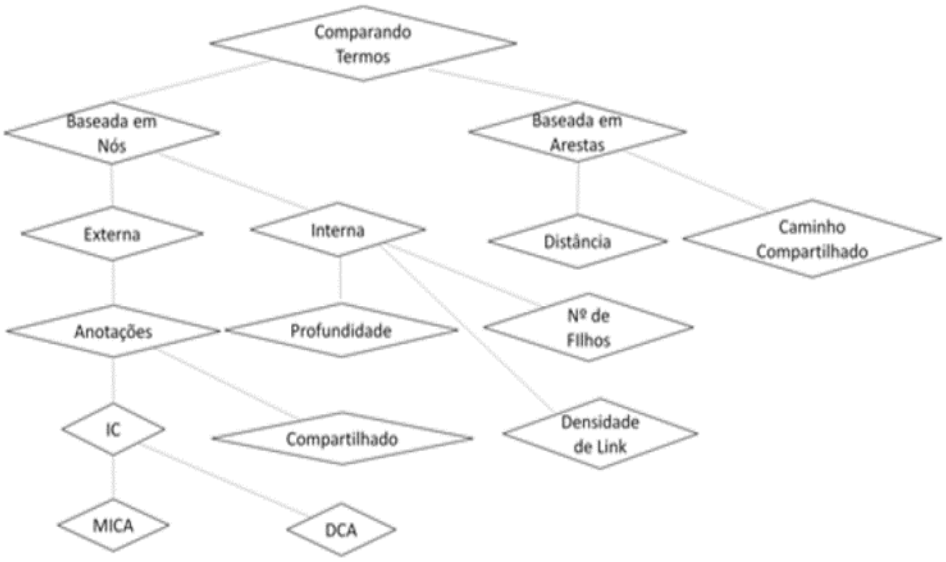
O termo “semântica”, em geral, trata do estudo do significado ou de ciência que estuda a significação (ALMEIDA, 2011). Os autores classificam a semântica considerando a forma como é expressa e para quem se destina sendo o processamento por máquinas, por exemplo, considerado como semântica explícita, especificada formalmente, para execução direta via mecanismos de inferências.

Segundo Gelaim (2013), mensurar a similaridade entre conceitos é possível através da consideração das suas características ou da distância entre eles, ou seja, quanto mais elementos em comum, mais similares.

A similaridade semântica baseada em estruturas de representação do conhecimento compara o significado, baseado nos conceitos que os termos representam e nos diferentes níveis de granularidade e abstrações.

Há uma série de abordagens disponíveis para quantificar a similaridade semântica entre termos ou entidades em uma ontologia representada em redes ou estruturas de grafos, tendo dois tipos de comparação sendo adotados com maior destaque: a baseada em arestas (*Edge-based*) e baseada em nós (*Node-based*), conforme Figura 5 (PESQUITA, 2009).

FIGURA 5 – Principais abordagens para comparação termos



Fonte: PESQUITA, 2009.

Abordagens baseadas em arestas baseiam-se principalmente na contagem do número de arestas no grafo caminho entre dois termos. A técnica mais comum calcula a distância que seleciona o caminho mais curto ou a média de todos os caminhos, quando existe mais do que um caminho. Esta técnica produz uma medida da distância entre dois termos que pode ser facilmente convertida em uma medida de similaridade. Uma forma para medir distância semântica é encontrada em representações taxonômicas, como árvores. Embora essas abordagens sejam intuitivas, eles são baseados em dois pressupostos que são raramente verdadeiros em ontologias biológicas: (1) nós e arestas são uniformemente distribuídas, e (2) arestas no mesmo nível na ontologia correspondem à mesma distância semântica entre os termos (PESQUITA, 2009).

Usualmente, esse tipo de abordagem se utiliza de relacionamentos do tipo “é-um” e “parte-todo” para definir relações de subclasses e superclasses entre os conceitos presentes na hierarquia das ontologias. O relacionamento semântico, neste caso, pode ser obtido usando o tamanho do caminho entre os termos (nos do grafo). “Um nó que tiver o menor caminho entre outro nó, é mais similar a ele” (RESNIK, 1999, p.96).

Abordagens baseadas em nós comparam as propriedades dos termos envolvidos que podem estar relacionados com eles, os seus antepassados ou seus descendentes. Um conceito comumente usado em todos é o conteúdo da informação (IC) que estabelece uma medida para valorar quanto específico e informativo é um termo.

Estudos sobre o desempenho das várias medidas de semelhança semântica têm revelado o uso do conteúdo de informação que dois conceitos partilham como uma técnica muito eficaz na comparação de conceitos (COUTO, 2011).

Na literatura há inúmeros modelos/algoritmos para cálculo de similaridade de conceitos, variando de acordo com sua aplicação. Segue abaixo exemplos de modelos para cálculo de similaridade:

O Modelo Relacional de Tversky (1977) avalia a similaridade considerando as características comuns e incomuns entre os estímulos (conceitos) e o contexto em que eles se encontram. Abaixo exemplo de cálculo representando através de lógica descritiva.

Sejam os conceitos:

Pai ≡Humano ∏ Masculino ∏∃ temFilho. Humano

Mulher ≡ Humano ∏ Feminino

A primeira parte do modelo consiste em extrair as características de cada conceito, assim, para Pai e Mulher são, respectivamente, {Humano, Masculino, ∃ temFilho.Humano} e {Humano, Feminino}. Em seguida separa-se o que é comum e incomum as operações de interseção e subtração da Teoria dos Conjuntos podem ser utilizadas, com isso:

Interseção:

{Humano, Masculino, ∃ temFilho.Humano} ∩ {Humano, Feminino} = {Humano}

Diferença:

{Humano, Masculino, ∃ temFilho.Humano} - {Humano, Feminino} = {Masculino, ∃ temFilho.Humano}

{Humano, Feminino} - {Humano, Masculino, ∃ temFilho.Humano} = {Feminino}

Portanto, os conjuntos obtidos com a interseção e diferença serão as entradas para as funções utilizadas no modelo de Tversky. A interseção é representada pela característica {Humano} e a diferença por {Masculino, ∃ temFilho.Humano}, para o pai e {Feminino} exclusiva de Mulher.

A segunda parte do modelo avalia o contexto que considera, por exemplo, que as características exclusivas do conceito {Pai} são mais relevantes do que {Mulher}, com isso o fator α representando o contexto de Pai será 2, e o fator β, das características do conceito Mulher será 0,5.

Com essas informações é possível avaliar a similaridade entre Pai e Mulher com a função do Modelo Relacional de Tversky (1).

$$s(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha * f(A - B) + \beta * f(B - A)} \qquad (1)$$

Sendo α, β ≤ 0. a e b os estímulos (conceitos), A e B conjuntos das características dos estímulos.

Mapeando o exemplo para a função, tem-se:

a = Pai;

b = Mulher;

A = {Humano, Masculino, ∃ temFilho.Humano}

B = {Humano, Feminino}

f(A∩B) = f({Humano});

f(A – B) = f({Masculino, ∃ temFilho.Humano });

f(B – A) = f({Feminino});

α = 2; β = 0,5;

Para calcular a similaridade é necessário definir a função f, assumindo que f seja a função de cardinalidade, então:

f(A∩B) = 1;

f(A – B) = 2;

f(B – A) = 1;

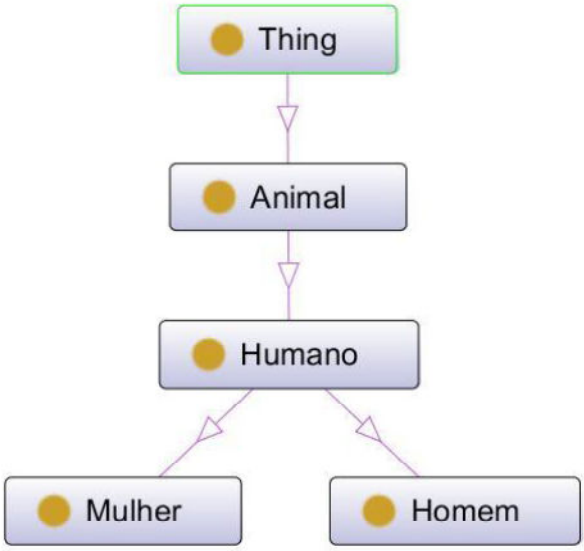
Com isso, a similaridade entre Pai e Mulher é dada por:

S(Pai, Mulher) = 1/(1 + 2*2 + 1*0,5) = 0,1818.

O modelo de rede semântica é uma notação gráfica para representar conhecimento sendo formada por nós (representam conceitos, objetos, propriedades) e arestas (relacionamento entre nodos) (GELAIM, 2013).

A lógica descritiva surgiu para prover significado declarativo e formal para as redes semânticas, preservando o formato estrutural para representação de conhecimento. A partir do relacionamento "é um" há várias propostas de avaliação da similaridade conceitual. Segundo o modelo proposto por Wu e Palmer (1994) *apud* (GELAIM, 2013), para medição de similaridade entre dois conceitos Conceito1 e Conceito2 são considerados o primeiro ancestral em comum (Conceito 3), o número N1 de nodos do caminho entre Conceito1 e Conceito 3, o número N2 de nodos do caminho entre Conceito 2 e Conceito 3 e o número N3 representando a distância entre Conceito 3 e a raiz.

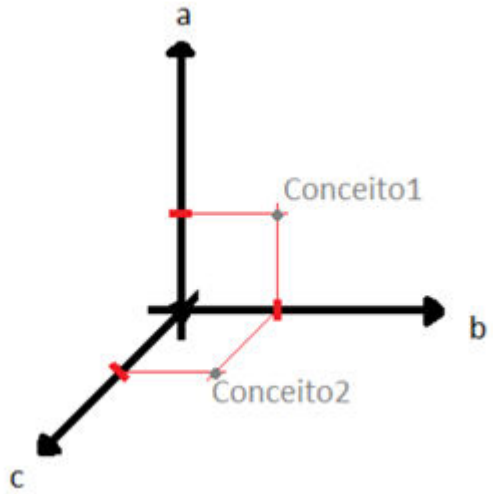
FIGURA 6 - Exemplo de estrutura conceitual



Fonte: GELAIM, 2013, p.43.

O modelo geométrico considera os conceitos como pontos no espaço, e a distância entre os pontos é a dissimilaridade entre eles, já que são as características não comuns que afastam os pontos (conceitos). Cada característica distinta do domínio é uma dimensão do espaço. A representação gráfica é dada na Figura 7. Com as coordenadas definidas calcula-se utilizando, por exemplo, a distância euclidiana.

FIGURA 7 – Representação da distância semântica



Fonte: GELAIM, 2013, p.42.

2.3 Ontologias e KDD

Segundo Ribeiro (2010) as ontologias podem trazer grandes benefícios praticamente para todas as fases do processo de KDD, seja auxiliando na escolha de algoritmos de MD, na elaboração de consultas mais eficientes, ou na visualização de novas regras de associação.

As ontologias são um modelo conceitual satisfatório para armazenar e manipular o conhecimento, resultando assim em uma melhor automação da Descoberta de Conhecimento, seja ela em qualquer uma das fases que constitui o processo de KDD (HAMANI, 2014).

2.3.1 Usos de ontologias em KDD

As ontologias estão sendo utilizadas para tornar mais relevantes os padrões descobertos pelas técnicas de mineração, principalmente, regras de associação e redes bayesianas.

Segundo Nigro *et al.* (2007), um dos problemas mais importantes e desafiadores do processo de KDD é a definição do conhecimento prévio; este pode ser originado a partir do processo ou do domínio. Esta informação do domínio pode ajudar a selecionar as informações, recursos ou técnicas adequadas, diminuir o espaço de hipóteses, representam a saída de uma forma mais compreensível e melhorar todo o processo. A ontologia pode representar o conhecimento do processo de descoberta de conhecimento e conhecimento sobre o domínio a ser explorado.

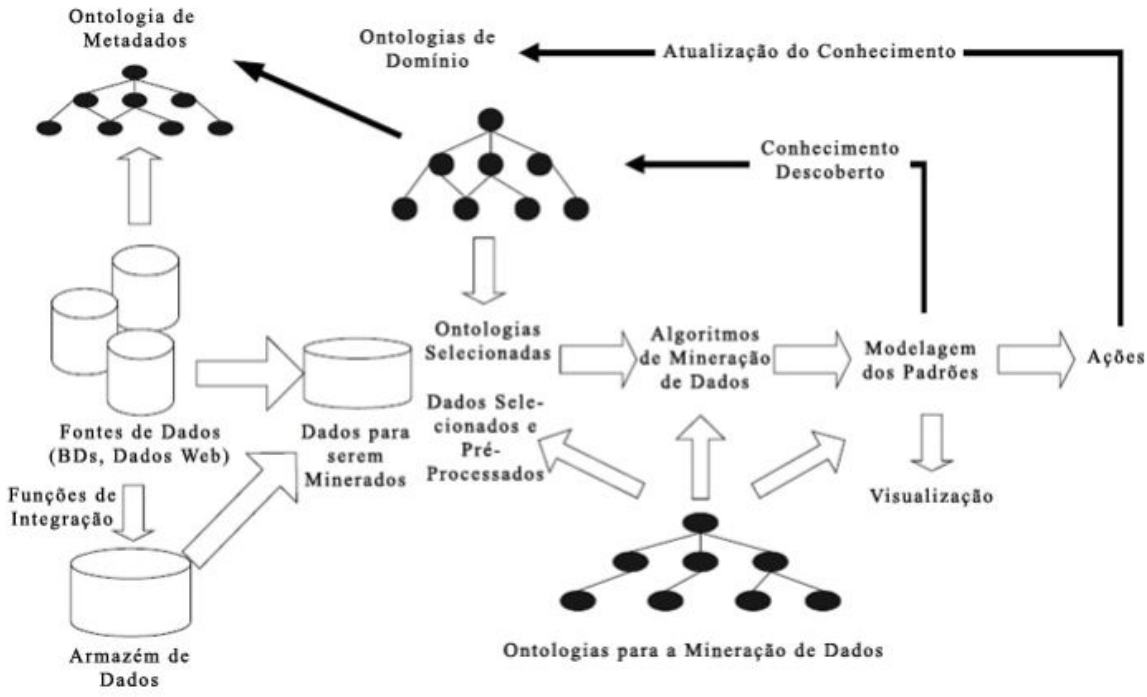
Segundo Nigro *et al.* (2007), as ontologias foram introduzidas na KDD pela primeira vez no começo de 2000 sendo utilizados, conforme Figura 8, de três formas:

- Ontologias para a mineração de dados – incorporar conhecimento ao processo de MD com a adição de ontologias para orientar o processo de descoberta, interpretação e validação do conhecimento extraído, ontologias para recursos e descrição do serviço;
- Mineração de dados para as ontologias – adicionar conhecimento de domínio à informação de entrada ou saída da MD. As aplicações mais características desta abordagem são em medicina, biologia e dados espaciais, como

a representação de genes, taxonomias, aplicações em geociências, aplicações médicas e especialmente na evolução de domínios;

- Ontologias de metadados - armazena o conhecimento sobre os atributos e seus relacionamentos extraídos da base de dados para compor uma ontologia que dá suporte a parte estrutural da base de dados.

FIGURA 8 - Framework geral para integração com mineração de dados



Fonte: NIGRO *et al.*,2007, p. xii.

O conteúdo semântico extraído das ontologias permite a inserção de mais inteligência e conhecimento em mineração de dados, melhorando sua qualidade (FERRAZ, 2008), pois permite a inserção de medidas subjetivas.

Nesse aspecto, a Ciência da Informação estrutura a aquisição da informação e do conhecimento na organização estabelecendo princípios e norteando a resolução do problema da representação do conhecimento com a adoção de ontologias para auxiliar o processo de KDD.

Boa parte da responsabilidade pelas dificuldades e fracassos em KDD é proveniente do desconhecimento sobre o negócio e dos dados submetidos à MD e à análise subsequente (GONÇALVES, 2004). Justamente por sua capacidade de tratar, incorporar e formalizar o conhecimento do negócio e do entendimento dos

dados, as ontologias possuem o potencial de auxiliar de forma decisiva na operacionalização e inserção do KDD em uma organização.

Ao integrar métodos objetivos com abordagem semântica é possível identificar com sucesso padrões não triviais, considerando o conhecimento do usuário (FERRAZ, 2008). Uso de Ontologias na mineração de dados por regras de associação.

A mineração de dados por regras de associação tradicional utiliza medidas objetivas estatísticas para obter o conjunto de itens frequentes, ou seja, desconsidera o significado de cada item ou instância. O conteúdo semântico extraído das ontologias permite a inserção de mais inteligência e conhecimento ao processo mineração de dados melhorando sua precisão (FERRAZ, 2008).

Segundo Ribeiro (2010) e Coelho (2012), o uso de ontologia pode reduzir os esforços dos especialistas de domínio na definição de regras de associação e na análise de descobertas de padrões. O conhecimento prévio de um domínio ou de um processo na área de mineração de dados pode ajudar a selecionar informações mais apropriadas (pré-processamento), diminuir o conjunto de dados a serem mineradas (processamento) e representar resultados de uma forma mais compreensível (pós-processamento) (HAMANI, 2014).

Segundo Ferraz (2008), as medidas de avaliação de regras interessantes consideram a relação entre o antecedente e o consequente. Diz-se que uma instância está coberta por uma regra, se o antecedente da regra for verdadeiro para a instância. Usualmente, esse tipo de abordagem se utiliza de relacionamentos do tipo “é um” (generalização e especialização) e “parte-todo” (composição) para definir relações de subclasses e superclasses entre os conceitos presentes na hierarquia da ontologia, ou seja, como balizadoras semânticas, que permitam uma redução do número de regras de associação, geradas pela mineração de dados, de maneira mais eficaz do que as tentativas meramente sintáticas, que permitam, ao mesmo tempo, um enriquecimento semântico do conjunto de regras mineradas.

As abordagens para poda e generalização de regras utilizam cálculo do grau de similaridade semântica entre termos que estão geralmente baseadas na estrutura hierarquia de grafos sendo, portanto, compatível com as ontologias que são representadas por grafos acíclicos orientados (DAG - *Directed Acyclic Graph*).

Segundo Ferraz (2008), a vantagem das restrições de poda é excluir a partir do início as informações que o usuário não está interessado. Cada regra mais geral deve ser capaz de substituir um número de regras específicas por meio de um processo de generalização. Quando isto é possível, há simultaneamente uma melhoria semântica do conjunto de regras de associação mineradas e uma futura redução na cardinalidade do conjunto de regras.

Segundo Nigro *et al.* (2007), através da representação do conhecimento por ontologia é possível transformar a MD em “mineração de conhecimento”.

2.4 Trabalhos relacionados

Alguns autores relatam em seus trabalhos como as ontologias estão sendo utilizadas para melhorar os resultados dos métodos de mineração de dados de regras de associação (FERRAZ, 2008; MARINICA, 2010; RIBEIRO, 2011; COELHO, 2012; VAVPETIC, 2012; MANDA, 2013; HAMANI, 2014).

Estes autores utilizam o tratamento semântico com ontologias de domínio para registro do conhecimento prévio e assim, na fase de pré-processamento ou pós-processamento, introduzir restrições dos usuários agrupadas em dois tipos: restrições de poda, para filtrar itens desinteressantes, e restrições de abstração, generalização de itens em relação a conceitos de ontologias.

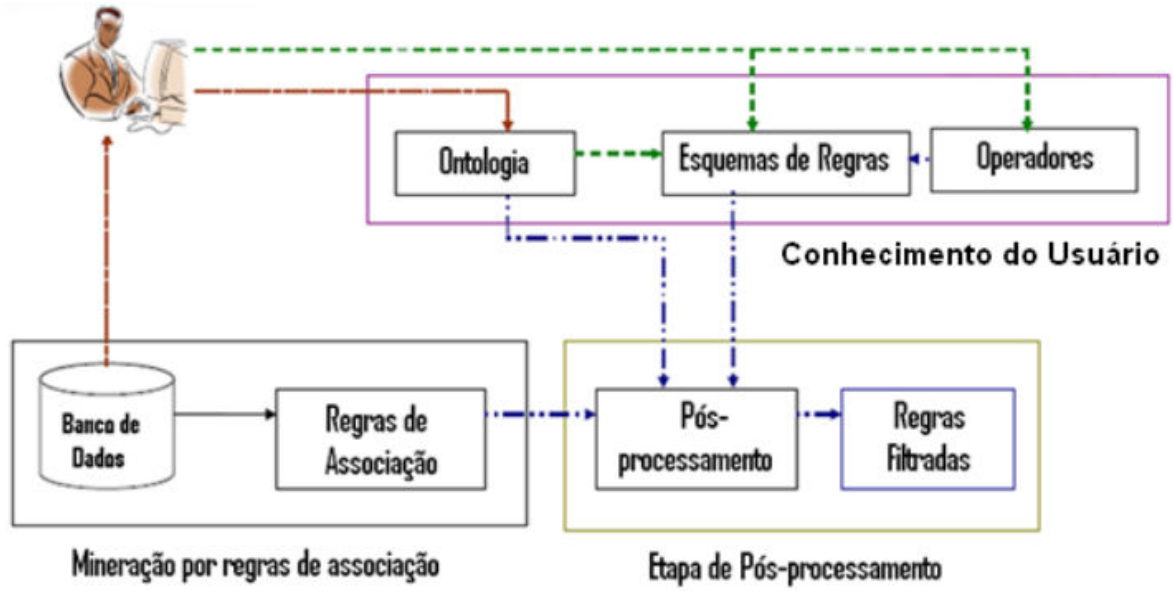
Tendo como base estes trabalhos buscou-se avaliar os métodos e técnicas a serem utilizados na elaboração do modelo de recuperação de informação a ser desenvolvido para a empresa de medicina diagnóstica na avaliação do comportamento de prescrição médica.

2.4.1 Aplicações da ontologia em mineração de dados por regras de associação

Coelho (2012) busca agregar o conhecimento de domínio através de ontologia difusa para propiciar suporte semântico na fase de preparação de dados e na fase de pós-processamento, auxiliando o analista a explicar e interpretar as regras obtidas. Com base nesta abordagem, nos testes e simulações realizados o autor conseguiu reduzir o número de regras de interesse em 42% com uma taxa de sucesso global de 95,88%. Na pré-mineração, com a preparação de dados a serem

apresentados ao minerador e na pós-mineração terá como domínio e escopo as regras obtidas no processo de MD para classificação das regras de associação obtidas, conforme Figura 9.

FIGURA 9 - Mineração de dados com suporte de ontologias



Fonte: COELHO, 2012, p. 77.

Marinica *et al.* (2009) utiliza ontologias para aprimorar a integração do conhecimento do usuário na tarefa de pós-processamento e esquema de regras (meta-regras) para especificar as expectativas do usuário.

No Quadro 2 são mostrados trabalhados utilizando ontologia e KDD.

QUADRO 2 - Trabalhos relacionando KDD e ontologia

Fase de KDD	Tipo de Aplicação da Ontologia	Exemplos
Integração Seleção	Ontologias como repositórios de conhecimento em relação ao domínio aplicado. É a fase de maior aplicação de Ontologias dentro do processo de KDD.	(SILVESCU *et al.,*2001), D2OMapper (ZU *et al*., 2006), SEMEDA (KÖHLER *et al.,* 2003), KAON (BOZSAK *et al*., 2002) , Escovar, Yaguinuma, Biaziz (2006), Chen *et al*. (2003), Brisson, Collard, Pasquier (2005), Panov *et al.* (2004). OntoDM
Limpeza dos Dados Redução dos	(1) ontologia para armazenar a informação necessária a transformar a instancia dos	(PHILLIPS; BUCHANAN, 2001), (BOWERS; LUDÄSCHER, 2004), (KEDAD ; METAIS, 2002),

Dados Transformação	dados; (2) ontologia para armazenar as instâncias mediante uma representação formal.	ONTOCLEAN (Wang *et al.,* 2005)
Mineração de Dados Interpretação	(1) ontologia utilizadas pelos analistas para uma melhor escolha do algoritmo de mineração, levando em conta os dados e informações disponíveis. Os resultados da análise dos dados também podem ser armazenados nas ontologias, e dessa forma a representação formal do novo conhecimento facilita sua reutilização; (2) ontologia de domínio para facilitar o estabelecimento de regras de associação e padrões durante a mineração de dados.	PROTEUS (CANNATARO *et al.,* 2005), IDEA (BERNSTEIN *et al*., 2005). Onto4KDD (GOTTGTROY *et al.,* 2004), LISp-Miner , (SVATEK *et al*., 2005), MiningMart (EULER; SCHOLZ, 2004) , Panavet (2008), OntoDM, Yokome (2011), Meta-DM, (COELHO, 2012), Ontologia com lógica difusa (MARINICA, 2009), Vanzin (2004), Ontologia com lógica difusa Brisson, Collard, Pasquier (2005) Vivacqua (2008), ontologia com cálculo de distância semântica (DS)

Fonte: Adaptado de RIBEIRO, 2010.

Manda *et al.* (2013) apresenta uma abordagem de mineração de dados baseada na utilização de multi-ontologia em todos os níveis (MOAL), que utiliza a estrutura e os relacionamentos de uma Ontologia Genética para minerar as regras de associação. O algoritmo consiste em três etapas principais:

1) Generaliza o conjunto de transações baseadas em relacionamentos transitivos, ou seja, infere a relação a partir de propriedades comum aos termos;

2) Avalia e classifica as regras utilizando métricas de interesse;

3) Poda as regras altamente relacionadas ou conhecidas considerando três tipos relações entre conceitos: hiperonímia (“é um”), meronímia (“parte de”) e de regulação (restrições).

Em alguns trabalhos, as medidas de semelhança semântica estão normalmente restritas às relações de hiperonímia e meronímia, por serem estas que definem os ancestrais e os descendentes de um dado conceito (COUTO, 2011).

Hamani (2014) aborda duas categorias principais de algoritmos para calcular a distância semântica entre os termos, organizados em estruturas hierárquicas, propostas na literatura:

1) Baseada na distância entre termos, ou seja, cálculo do caminho mais curto entre dois termos (nós do grafo).

2) Baseada no conteúdo da informação considerando a ideia que pares de palavras que compartilham muitos contextos comuns são semanticamente relacionados.

Manda (2013) acrescenta mais três categorias de métodos que utilizam ontologias para cálculo da similaridade semântica: métodos baseados em propriedades de termos, métodos baseados na hierarquia da ontologia e métodos híbridos. Para o autor quatro fatores são normalmente considerados em métodos baseados em distância como segue:

(1) Densidade da ontologia: maior será a densidade quanto mais perto a distância entre os nós;

(2) Profundezas de nós: quanto mais profundos os nós localizados é mais evidente a diferença entre os nós;

(3) Tipos de ligações: o tipo normal “é um”, e outras relações, tais como “parte-todo” são associados pesos diferentes entre os nós;

(4) Pesos de links: pesos distintos para arestas que conectam um certo nó com todos os nós filhos.

Através dos trabalhos desenvolvidos, observa-se que a associação entre KDD e ontologia é fundamental para agregar conhecimento e valor ao processo.

3 METODOLOGIA

A abordagem proposta consiste na hipótese de que os dados contidos nas bases de dados possuem uma relação com os conceitos envolvidos no negócio e os relacionamentos existentes entre eles o que conduzirá a descoberta de semântica implícita conduzindo à seleção dos subconjuntos de dados que devem ser minerados.

O presente capítulo apresenta a metodologia empregada desde a concepção da ontologia até o processo de KDD.

Baseada na metodologia de pesquisa científica de Lakatos e Marconi (1991) e Gil (1994), a pesquisa pode ser classificada:

• Quanto à natureza, como pesquisa aplicada, pois objetiva gerar resultados para aplicação prática pelas empresas de medicina diagnóstica;

• Quanto a abordagem do problema, pesquisa qualitativa e quantitativa, visto que pretende validar a ontologia com os especialistas através de entrevistas e questionários e obter padrões através da extração em base de dados utilizando técnicas estatísticas.

• Do ponto de vista dos objetivos, como pesquisa explicativa, pois pretende ilustrar como é possível aprimorar o processo de descoberta de conhecimento em base de dados com o uso de ontologias;

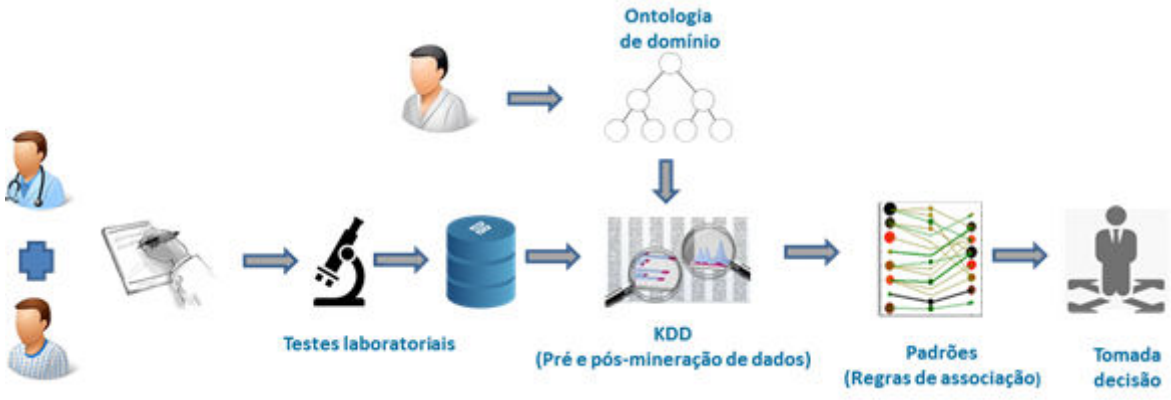
• Considerando os procedimentos técnicos, como pesquisa experimental, pois busca comprovar que o modelo proposto pode melhorar a recuperação de informação.

Segundo Vidigal (2011), a metodologia qualitativa trabalha com valores, crenças, representações, hábitos, atitudes e opiniões e se aplica em processos particulares de grupo mais específicos e delimitados. Uma boa pesquisa qualitativa exige que as fontes tenham conhecimento sobre o assunto e liberdade para falar sobre o mesmo. Para tanto, foi realizado entrevista com os especialistas, conforme apêndice A, sendo adotada uma amostra não probabilística por conveniência, onde os entrevistados foram selecionados por serem especialistas (biomédicos, farmacêuticos ou médicos) em uma empresa de medicina diagnóstica.

Assim, o presente estudo buscou recuperar o conhecimento sobre as hepatites virais através da representação ontológica a fim de disponibilizar

conhecimento necessário que subsidie o processo recuperação da informação para a tomada de decisão, conforme mostrado na Figura 10.

FIGURA 10 – KDD com ontologias de domínio



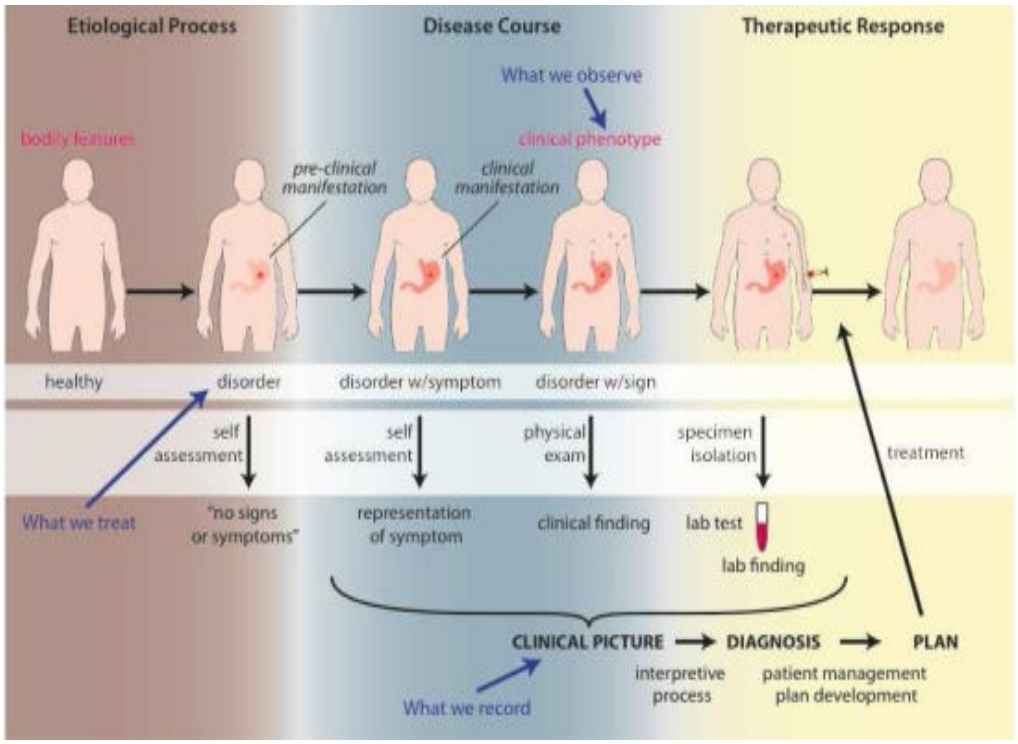
Fonte: Elaborado pelo próprio autor.

O diagnóstico médico é o processo avaliação do paciente para identificação de uma dada doença que tem como base a utilização de protocolos que emerge sob a forma de um quadro descritivo de sinais e sintomas, alterações anatomopatológicas que lhe seja reconhecido e na procura dos agentes etiológicos das lesões.

O processo de diagnóstico diferencial, a escolha dos exames complementares e a interpretação de resultados são habilidades fundamentais para o médico que utiliza das informações obtidas na anamnese e exame clínico para analisar os dados preliminares e direcionar a causa mais provável do problema do paciente (STERN *et al.*,2007).

Scheuermann et al. (2009), mostra na Figura 11 como se manifesta a doença e como ocorre o processo de diagnóstico. Uma pessoa saudável ainda sem sinais ou sintomas tem alterações no organismo. Posteriormente, os sinais e sintomas se manifestam. O paciente procura o médico que realiza o exame físico e testes laboratoriais (exames complementares) de amostras derivadas do paciente, cujos resultados podem ser gravados no registro médico como um quadro clínico. O quadro clínico é interpretado pelo médico para se chegar a um diagnóstico, que serve por sua vez como a base para o desenvolvimento de um plano tratamento do paciente.

FIGURA 11 - Quadro terminológico de diagnóstico de uma doença



Fonte: SCHEUERMANN *et al.,* 2009, p.120.

Para integrar estes dados, conflitantes ou pouco confiáveis, obtidos com a anamnese e exame clínico e obter um possível diagnóstico da doença apresentada pelo paciente é necessário elaborar uma lista de possíveis causas e priorizar.

Para elaborar uma lista de possíveis causas são utilizadas técnicas de memorização, arcabouços anatômicos, segundo órgãos/sistemas, fisiopatológico, mnemônicos ou a combinação de várias técnicas. Já a priorização é possível com adoção das abordagens, separadas ou conjuntamente, a saber (STERN *et al.*,2007):

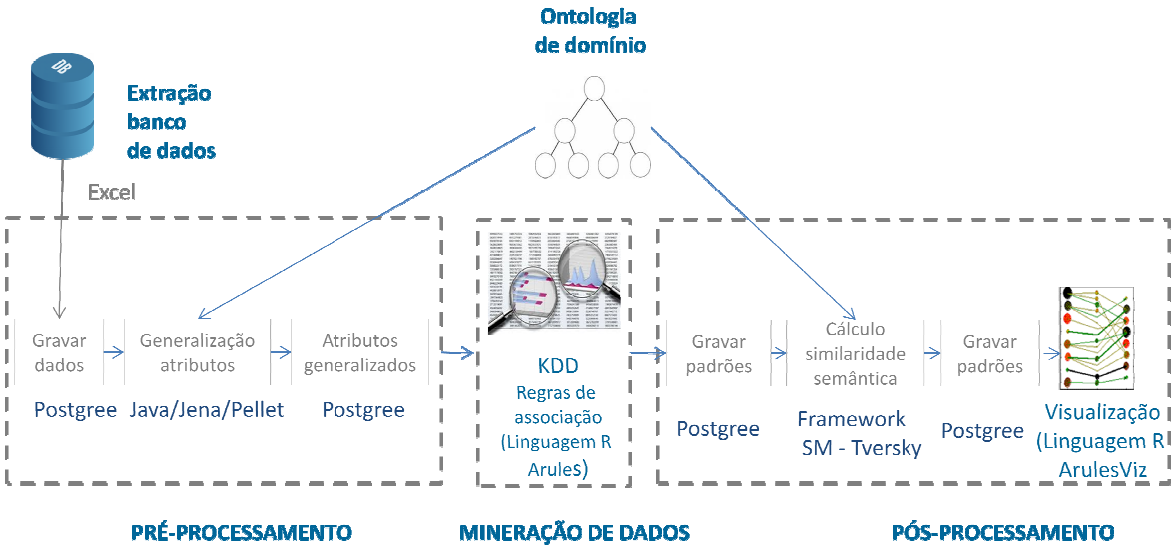
- Possibilística - testa simultaneamente todas as causas conhecidas;
- Probabilística - considera os transtornos mais prováveis;
- Prognostica - considera os diagnósticos mais sérios;
- Pragmática - considera os diagnósticos mais responsivos ao tratamento.

Após elaborar as hipóteses para as possíveis doenças é necessário avaliar se há necessidade de informações adicionais antes de prosseguir com o tratamento ou descartar o diagnóstico. Se a hipótese principal e as alternativas não forem válidas prossegue com o processo de diagnóstico buscando novas causas.

Portanto, quando apenas a avaliação clínica não é suficiente, a realização de exames complementares é utilizada para confirmar ou descartar a doença. Neste momento, a medicina diagnóstica é acionada para realização dos exames complementares necessários para confirmação do diagnóstico e tratamento.

Com a base de testes complementares é possível através da utilização da técnica de KDD por regras de associação buscar padrões de comportamento do prescritor fazendo com que o tomador de decisão possa direcionar as ações de divulgação e venda. O modelo da figura 12 representa a solução tecnológica utilizada nesta pesquisa.

FIGURA 12 - Framework da solução



Fonte: Elaborado pelo próprio autor.

3.1 Contexto

Nessa seção apresenta-se um panorama do segmento de Medicina Diagnóstica no qual a pesquisa é aplicada.

O mercado de saúde no Brasil é da faixa de $270 bilhões sendo 54% para o setor público e 46% de investimento privado. O investimento público atende 74% da população e o restante é atendido pelo privado. Na medicina diagnóstica são

investidos 3% dos gastos públicos com saúde; já o investimento privado, direcionada 7% para esta área (Formato Clínico, 2012).

Ao longo dos últimos anos, o mercado de saúde tem sofrido muitas transformações. Na tentativa de oferecer serviços de saúde pública com um menor custo e maior eficiência, o governo tem aumentado o número e os tipos de serviços terceirizados, prestados por companhias privadas. Em vista disso, a Agência Nacional de Saúde Suplementar (ANS) foi criada por ocasião da regulamentação do setor de saúde suplementar, a partir da Lei nº 9.656/98, e tem por finalidade institucional promover a defesa do interesse público na assistência suplementar à saúde, regular as operadoras setoriais - inclusive quanto às suas relações com prestadores e consumidores - e contribuir para o desenvolvimento das ações de saúde no país.

Devido este panorama e da inexistência de divulgação de dados consolidados sobre o mercado de medicina diagnóstica atendida pela saúde suplementar, a empresa em questão busca aprimorar seu monitoramento ambiental para melhor entender a real demanda por serviços neste segmento direcionando suas ações comerciais e de marketing. A recuperação de informação sobre o mercado utilizando análise da base dados da empresa é uma excelente oportunidade para extrair conhecimento.

3.1.1 A instituição objeto da pesquisa

A empresa estudada é uma organização nacional fundada nos anos 50 que atua no segmento de medicina diagnóstica. A medicina diagnóstica contempla diferentes especialidades direcionadas à realização de exames complementares no auxílio ao diagnóstico médico.

A empresa tem atuação nacional na prestação de serviços de análises clínicas, criopreservação e genética no segmento de apoio a outros laboratórios e em suas unidades de captação própria, também, oferece serviços de vacinas, diagnóstico por imagem e *check up*.

A empresa, nos últimos anos, realiza uma série de ações com o intuito de aprimorar o conhecimento sobre o mercado, agregando valor para os clientes e tornando a empresa referência nacional em medicina diagnóstica e preventiva. A

área de inteligência competitiva é responsável pelo levantamento dos dados necessários para revisão anual do planejamento estratégico e pelo levantamento das informações para tomada de decisão pela alta direção e gerências.

Ao obter informações mais interessantes o especialista KDD poderá ajudar na geração de conhecimento pelas equipes de Inteligência competitiva, comercial, produtos, novos negócios e de marketing na segmentação dos produtos no mercado.

3.1.2 Condução e recorte da pesquisa

Atualmente há disponíveis no mercado de medicina diagnóstica inúmeros exames complementares para auxílio ao diagnóstico de doenças. Segundo Capilheira e Santos (2006) a prescrição de exame complementar é influenciada por fatores relacionados à: organização do sistema de saúde, características do médico e do paciente, além das diferentes interações entre esses grupos.

Optou-se por analisar a prescrição médica dos testes laboratoriais relacionados ao diagnóstico de hepatites virais limitando o universo de pesquisa, já que, há um grande número de testes laboratoriais disponíveis no mercado e sua identificação é complexa, pois está diretamente relacionado ao diagnóstico médico o qual necessita de um conhecimento especializado.

As hepatites virais estão incluídas na lista de doenças de notificação compulsória e, portanto, os profissionais de saúde têm papel relevante na notificação e no acompanhamento das pessoas portadoras, sintomáticas ou não. Para que possam exercer tal papel, é necessário que esses profissionais estejam aptos a identificar casos suspeitos, solicitar exames complementares adequados e realizar o encaminhamento dos casos indicados a serviços de referência.

As hepatites virais são doenças provocadas por diferentes agentes etiológicos, com tropismo (propensão que um vírus tem em infectar determinado tipo de célula ou tecido em especial) primário pelo fígado, e apresentam características epidemiológicas, clínicas e laboratoriais distintas. A distribuição das hepatites virais é

universal, sendo que a magnitude varia de região para região, de acordo com os diferentes agentes etiológicos.

As hepatites virais têm grande importância para a saúde pública e para o indivíduo, pelo número de indivíduos atingidos e pela possibilidade de complicações das formas agudas e crônicas. O homem é o único reservatório com importância epidemiológica. Os outros reservatórios apresentam importância como modelos experimentais para a pesquisa básica em hepatites virais.

Os agentes etiológicos que causam hepatites virais mais relevantes do ponto de vista clínico e epidemiológico são designados por letras do alfabeto (vírus A, vírus B, vírus C, vírus D e vírus E). Estes vírus têm em comum a predileção para infectar os hepatócitos (células hepáticas). Entretanto, divergem quanto às formas de transmissão e consequências clínicas advindas da infecção. São designados rotineiramente pelas seguintes siglas: vírus da hepatite A (HAV), vírus da hepatite B (HBV), vírus da hepatite C (HCV), vírus da hepatite D (HDV) e vírus da hepatite E (HEV), conforme Figura 13. A doença tem um amplo espectro clínico, que varia desde formas assintomáticas, ictéricas típicas, até a insuficiência hepática aguda grave (fulminante).

Existem outros vírus que podem causar hepatite (ex: TTV, vírus G, SEV-V), mas seu impacto clínico e epidemiológico é menor estando basicamente concentrado em centros de pesquisa.

FIGURA 13 - Principais características dos vírus que causam a hepatite

Agente etiológico	Genoma	Modo de transmissão	Período de incubação	Período de transmissibilidade
HAV	RNA	Fecal-oral	15-45 dias (média de 30 dias)	Desde duas semanas antes do início dos sintomas até o final da segunda semana da doença
HBV	DNA	Sexual, parenteral, percutânea, vertical	30-180 dias (média de 60 a 90 dias)	Duas a três semanas antes dos primeiros sintomas, se mantendo durante a evolução clínica da doença. O portador crônico pode transmitir o HBV durante anos
HCV	RNA	Parenteral, percutânea, vertical, sexual	15-150 dias	Uma semana antes do início dos sintomas e mantém-se enquanto o paciente apresentar HCV-RNA detectável
HDV	RNA	Sexual, parenteral, percutânea, vertical	30-180 dias. Este período é menor na superinfecção	Uma semana antes do início dos sintomas da infecção conjunta (HBV e HDV). Na superinfecção não se conhece este período
HEV	RNA	Fecal-oral	14-60 dias (média de 42 dias)	Duas semanas antes do início dos sintomas até o final da segunda semana da doença

Fonte: BRASIL.Secretaria de Vigilância em Saúde / MS, 2010, p. 411.

3.2 Modelagem ontológica

Neste capítulo, serão descritos os passos seguidos para a construção do sistema de organização do conhecimento, na forma de ontologia, sobre os exames complementares para diagnóstico de hepatites virais.

3.2.1 Propósito

A demanda por modelagem do conhecimento de forma clara e sem ambiguidade é um requisito imprescindível para aprimorar o processo de recuperação de informação trazendo maior precisão e relevância para os usuários. Sendo assim, as ontologias foram propostas como um modo de representar conhecimento e possibilitar a integração de informações.

A presente ontologia, detalhada no apêndice C, criada e disponibilizada como Ontologia de testes laboratoriais para Hepatite Viral (HVO), propõe-se a representar um domínio especifico para exames complementares relacionados ao diagnóstico de hepatites virais sendo desenvolvida usando a linguagem OWL, especificamente a OWL-DL.

Os usuários de tal artefato serão profissionais de saúde (médicos, farmacêutica biomédicos) e os profissionais de tecnologia de informação que desenvolvem sistemas e aplicações na área da saúde para uso pelas áreas de inteligência competitiva, comercial, *marketing*, produtos entre outras.

3.2.2 Escopo

O objetivo da ontologia HVO – ontologia de testes laboratoriais das hepatites virais é descrever os exames complementares relacionados ao diagnóstico das hepatites virais, tais como, utilizados para triagem da doença, no acompanhamento da evolução clínica e outros que não estão relacionados diretamente ao diagnóstico da doença, mas são rotinas para acompanhamento do paciente.

A ontologia aqui descrita, a HVO, apresenta uma estrutura hierárquica concebida a partir da *Ontology for General Medical Science* (OGMS). A OGMS se

fundamenta nos princípios da *Basic Formal Ontology* (BFO) (GRENON *et al*., 2004), uma ontologia de alto-nível criada em 2002 para apoiar pesquisas científicas elaborada utilizando guia e melhores prática da *Open Biomedical Ontology* (OBO) *Foundry ontologies*.

Os principais componentes a serem descritos na HVO são relacionados à representação dos testes laboratoriais e a descrição temporal de processos que são vinculados ao diagnóstico das hepatites virais nos indivíduos. Entre estes está à representação dos processos de solicitação de testes laboratoriais além de outros fatores os quais, sendo realizados, podem contribuir para avaliação clínica do paciente.

Considerando os objetivos supracitados, foi adotada para o desenvolvimento da ontologia de exames complementares para diagnóstico das hepatites virais a metodologia *Ontology development* 101, que trata do desenvolvimento de ontologias através da adoção de sete passos: (a) determinar o domínio e escopo da ontologia; (b) considerar o reuso de ontologias existentes; (c) enumerar termos importantes na ontologia; (d) definir as classes e a hierarquia das classes; (e) definir as propriedades das classes; (f) definir as restrições e por fim, (g) criar instâncias.

3.2.3 Fontes de informação

As fontes de conhecimento utilizadas, os termos empregados e suas respectivas descrições são provenientes de publicações relacionadas ao diagnóstico de hepatites virais (livros, periódicos nacionais/internacionais, LOINC, SNOMED-CT, e do Ministério da Saúde brasileiro, entre outros). Vale ressaltar que grande parte das entidades descritas na HVO são provenientes da OGMS e algumas da IDO[13], DOID, OBI [14] e FMA. Adicionalmente, outras ontologias e suas respectivas descrições sobre o assunto foram verificadas para esclarecer dúvidas.

Como referência dos testes laboratoriais existentes no mercado para diagnóstico das hepatites virais foi adotado o *Logical Observation Identifiers Names*

[13] Disponível em:<http://infectiousdiseaseontology.org/page/Main_Page>. Acesso em: 1 mai. 2014.

[14] Disponível em:<http://obi-ontology.org/page/Main_Page>. Acesso em: 1 mai. 2014.

and Codes (LOINC) através de pesquisa utilizando a ferramenta *Regenstrief LOINC Mapping Assistant* (RELMA). Outra fonte importante de conhecimento para o presente trabalho são as publicações do ministério as saúde sobre os protocolos de atendimento quando da suspeita de infecção por hepatites virais.

Seguindo o modelo proposto por Coelho (COELHO; ALMEIDA, 2012), a aquisição de conhecimento sobre o domínio foi conduzida através de um levantamento inicial com a pesquisa de dados secundários. Em um segundo momento foi realizado o roteiro da entrevista com os especialistas (Apêndice A) que objetivou a coleta de informações para se obter respostas para validar o conhecimento existente e esclarecer dúvidas e mal entendidos.

Após a realização das entrevistas, de acordo com o perfil da pesquisa, foram seguidas as etapas descritas abaixo, conforme abordagem adotada por Vidigal (2011):

- Leitura das entrevistas transcritas de forma sistemática e interativa;
- Identificação dos elementos comuns e divergentes;
- Organização e categorização do material;
- Reorganização do material em torno do tema e dos objetivos da pesquisa;
- Tratamento e análise do material;
- Elaboração do texto final.

3.2.4 Integração com outras ontologias

A HVO, para uma representação mais completa do domínio de testes laboratoriais para hepatites virais, importa e estende a OGMS, pois considera o processo de diagnóstico médico como ponto de partida para a prescrição dos testes complementares.

Devido o universo de testes complementares está diretamente relacionada com a doença fez-se necessário avaliar outras ontologias biomédicas existentes no *Open Biomedical Ontologies Foundry* [15] sendo considerada a possibilidade de reuso das ontologias descritas abaixo:

[15] Disponível em:< http://www.obofoundry.org/>. Acesso em: 1 mai. 2014.

- *Infectious Disease Ontology* (IDO) - Trata-se de um conjunto de ontologias do domínio das doenças infecciosas;
- *Ontology for General Medical Science* (OGMS) - Ontologia que aborda questões relacionadas com o diagnóstico e tratamento de doenças;
- *Disease Ontology* (DOID) - Ontologia de código aberto para a integração de dados biomédicos associada com a doença humana;
- *Ontology for Biomedical Investigations* (OBI) - Ontologia desenvolvida para integração de descrição de investigação clínica e biológica;
- *Foundation Model of Anatomy ontology* (FMA) – Ontologia para representação de classes ou tipos e relações necessárias para a representação simbólica da estrutura fenotípica do corpo humano.

Devido à importância das terminologias no contexto da organização da informação na área médica, foi utilizado como apoio na construção da ontologia de medicina diagnóstica o *Systematized Nomenclature of Medicine-Clinical Terms* (SNOMED-CT) (SCHULZ, JANSEN, 2013; SMITH *et al.,* 2007).

3.2.5 Desenvolvimento da ontologia

Do ponto de vista da modelagem da ontologia em uma linguagem formal, compreensível e processável por máquina, a HVO foi concebida sobre uma ótica formal, sendo modelada em OWL DL (BAADER *et al*., 2003).

A ontologia de aplicação construída ser propõe a resolver questões relacionadas ao processo de KDD em uma empresa de diagnóstico laboratorial, portanto, não pretende ser considerada uma ontologia de referência para este domínio.

Não há uma metodologia padrão para construção de ontologias, portanto, cada desenvolvedor usa aquela que se relaciona melhor com seus próprios métodos e critérios. Na construção da ontologia HVO foi utilizado o método *Ontology 101* (NOY; MCGUINNESS, 2001). A escolha desta metodologia se deve aos relatos da simplicidade na sua adoção e a adaptação com a ferramenta Protégé.

3.2.6 Softwares utilizados

Para modelagem da ontologia foi utilizado o editor de ontologias Protegé v.5[16], com motor de raciocínio Pellet v.2.3.0[17] para classificação automática de classes e instancias. Com esta modelagem compreensível e processável por máquina foi possível obter um arquivo OWL salvo pelo Protege no formato RDF o qual será posteriormente utilizado para generalização dos atributos a serem minerados.

3.3 KDD

A abordagem desta pesquisa baseia-se na aplicação de algoritmos de mineração de dados nos registros obtidos após generalização dos atributos com o uso de uma ontologia de domínio. A técnica de modelagem escolhida foi a geração de regras de associação, por meio do algoritmo Apriori, captando assim o perfil de prescrição médica. Como as solicitações envolvem o diagnóstico de várias doenças foram selecionados apenas os pedidos que possuíam algum exame diretamente relacionado ao diagnóstico de hepatites virais.

Para exploração, na busca de padrões relevantes, uma amostra representativa da base de dados foi extraída do total de exames disponíveis.

Para dar suporte aos procedimentos realizados neste trabalho, optou-se por seguir um modelo de processo de KDD conhecido como CRISP-DM.

De acordo com o modelo CRISP-DM o processo consiste em seis fases: compreensão do domínio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição dos resultados obtidos (CHAPMAN *et al*., 2000).

3.3.1 Propósito

O contexto do processo está relacionado ao levantamento da relação entre os testes laboratoriais prescritos pelos médicos para diagnóstico de hepatites virais.

---

[16] Disponível em:<protegewiki.stanford.edu>. Acesso em: 1 .mar. 2014.

[17] Disponível em:<http://clarkparsia.com/pellet>. Acesso em: 1 .fev. 2015.

O problema de mineração de dados em questão envolve a geração de regras de associação entre estes testes laboratoriais.

A estrutura dos dados das solicitações de pedido médico está armazenada em um banco de dados relacional, onde foram efetuadas transformações em sua estrutura. As transformações utilizadas são descritas, em detalhe, na fase de preparação dos dados. Após a geração das regras, estas foram armazenadas em uma tabela no Postgree de forma a facilitar análise de similaridade semântica entre os conceitos.

3.3.2 Compreensão do domínio

A fase de compreensão do domínio contemplou desde a aquisição do conhecimento para elaboração da ontologia HVO até as pesquisas para uso da técnica de mineração por regras de associação possibilitando que a extração dos dados da base de dados fossem mais assertiva.

3.3.3 Entendimento dos dados

O conhecimento prévio sobre o domínio e a aquisição de conhecimento para elaboração da ontologia facilitaram o entendimento dos dados.

Com o mapeamento prévio na ontologia das propriedades de anotação das classes foi identificado quais os exames complementares são diretamente associados ao diagnóstico das hepatites virais. Com os códigos de identificação foram extraídos da base de dados da empresa de medicina diagnóstica apenas os pedidos que continham alguns destes exames complementares.

3.3.4 Coleta de dados e descrição

No levantamento dos dados foram extraídos da base de dados os pedidos e seus respectivos testes laboratoriais. Estes pedidos foram separados em duas linhas de análise: pedidos cadastrados nas próprias unidades da empresa de medicina diagnóstica e pedidos encaminhados por laboratórios parceiros. Esta separação é importante, pois na segunda amostra, testes laboratoriais inespecíficos, ou seja, de rotina, normalmente, não estão presentes, pois são realizados internamente pelos laboratórios parceiros.

Assim, o conjunto de dados selecionado foi filtrado para manter apenas os pedidos que possuíam algum teste laboratorial diretamente relacionado ao diagnóstico das hepatites virais.

O conjunto de dados foi armazenado no banco de dados PostgreSQL 9.4 contendo quatro tabelas*: pedido, exame, pedido_exame, pedido_inferencia e exames_não_inferidos,* conforme apresentado na figura 13. Na tabela *pedido* está armazenado as informações relativas ao paciente no qual o exame foi realizado. Sempre que uma amostra biológica é coletada para realização de testes laboratoriais algumas informações para identificação do cliente são obrigatoriamente solicitadas. Para este trabalho foram extraídos apenas alguns atributos do pedido: unidade (identificador da unidade ou laboratório parceiro onde a coleta foi realizada), pedido (identificador único da amostra biológica coletada), dtped (data do pedido), *nomeCliente* (nome do cliente cuja amostra biológica foi coletada), idade (idade do cliente), sexo (sexo do cliente), estado (estado do cliente). Cada linha na tabela pedido representa um pedido médico para um cliente.

Na tabela *pedido_exame* são armazenados os testes laboratoriais solicitados. Além da informação do pedido constam o código do exame complementar solicitado. Um pedido pode ter vários exames complementares.

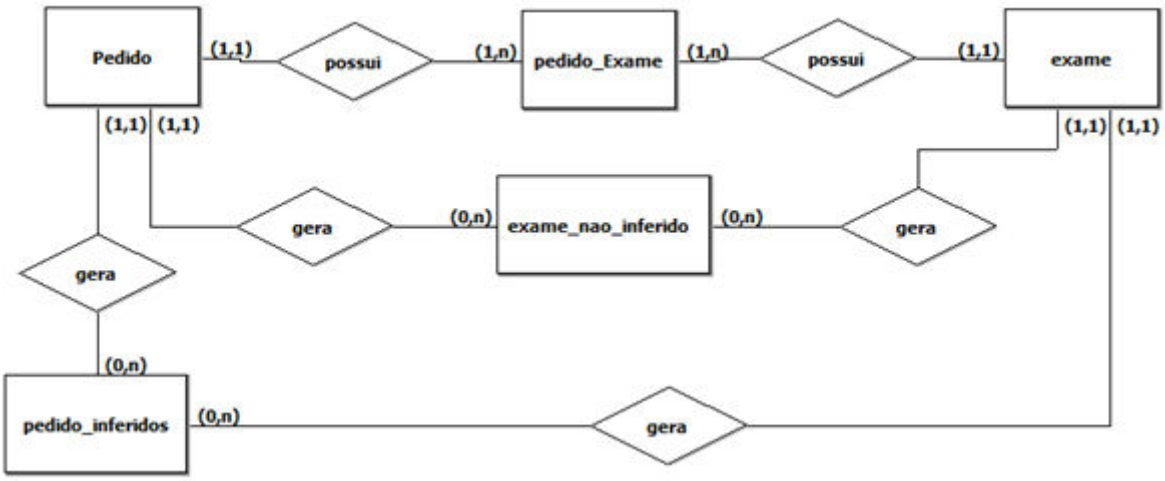
Um mesmo cliente pode aparecer em mais de uma linha na tabela de pedido. No entanto, para este trabalho não está sendo avaliado cada paciente individualmente, mas o comportamento de prescrição médica.

Na tabela *pedido_inferido* são armazenados o pedido, a identificação da doença que está sendo avaliado e os exames considerados na inferência gerados a partir da relação de exames validados através do axioma de equivalência, exemplificado na figura 18.

Já na tabela *exames não_inferidos* constam o pedido e o código dos exames que não estão relacionados diretamente ao diagnóstico das hepatites virais.

A tabela *exames* contempla a identificação e descrição dos testes complementares.

FIGURA 14 – Modelo relacional



Fonte: Elaborado pelo próprio autor.

3.3.5 Preparação dos dados

Após as fases de compreensão do domínio e entendimento dos dados, foi realizada a fase de preparação dos dados. O objetivo principal dessa fase foi a seleção, seguida da preparação dos dados armazenados no sistema de gerenciamento de banco de dados, para um formato adequado a aplicação do algoritmo Apriori (AGRAWAL *et al.,* 1993), utilizando a linguagem R.

A tabela *pedido* possui 34411 linhas, que representam uma amostra de 70% do número de pedidos distintos que foram realizados, no período compreendido entre janeiro de 2015 e março de 2015. A tabela *pedido_exame* possui 168800 linhas, onde cada linha contém a informação de cada teste laboratorial solicitado para o pedido. É importante notar que pedido representa uma solicitação médica de exames complementares ao diagnóstico clínico, logo, o mesmo cliente pode aparecer em vários pedidos, pois este pode ter comparecido no laboratório em momentos distintos. Em média um pedido contém 5 exames complementares. No entanto, o total de exames distintos nesta amostra é de 465.

Os dados disponíveis nas tabelas *pedido* e *pedido_exame* no banco de dados foram tratados antecipadamente, portanto, não houve necessidade de aplicar procedimentos para limpeza dos dados e complementação de dados faltantes.

Para fins desta pesquisa será utilizado apenas os atributos dos pedidos e respectivos exames, pois o intuito deste trabalho é mostrar as regras de associação entre eles independente de sexo, idade e localização do paciente.

A base de dados dos pedidos cadastrados nas unidades, denominada Base Unidades, e dos laboratórios parceiros, denominada Base Terceirizados, foram tratadas separadamente, já que a segunda contém exames que normalmente são terceirizados, portanto, excluindo exames de rotina.

Os dados foram extraídos da base da empresa de medicina diagnóstica considerando os pedidos que continham exames complementares específicos para diagnóstico de hepatites virais, conforme levantamento modelado na ontologia. No Quadro 3 estão representados alguns pedidos. Os dados obtidos foram processados pelo pacote ARules[18] da linguagem R.

QUADRO 3 – Relação de pedidos de exames – Unidades

Ped	DtPed	Idade	Sexo	Exames
8000001	01/01/2015	37ª	F	UR
8000001	01/01/2015	37ª	F	HBC-G
8000001	01/01/2015	37ª	F	|HBC-M
8000028	01/01/2015	42ª	F	AU|CBI|CHLA-G|CHLA-M|E2|FSH|GS-HTLV1
8000028	01/01/2015	42ª	F	AU
8000028	01/01/2015	42ª	F	ELISAG
8000101	02/01/2015	49ª	M	HCV

Fonte: Elaborado pelo próprio autor.

O algoritmo *Apriori* encontra os grupos de itens ocorrendo frequentemente juntos em transações. Nesse contexto, esses grupos são exames complementares, juntos, em um mesmo pedido. Esses conjuntos de itens são denominados conjuntos frequentes de itens. Conjuntos frequentes de itens são gerados com base no suporte fornecido como entrada no algoritmo. Assim é importante definir um valor apropriado de suporte, tal que o algoritmo possa encontrar padrões relevantes. Como a medida de suporte corresponde a avaliação de frequência com A e B em toda a base foi necessário adotar valores de suporte distintos por base analisada para que itens mais relevantes não fossem descartados.

[18] Disponível em:< https://cran.r-project.org/web/packages/arulesViz/index.html>. Acesso em: 10.mai.2014.

Foi utilizado com valor de suporte para Base Unidades e Base Terceirizados, respectivamente, sup=0,2 e sup=0,01 e confiança de 75%. As regras geradas foram armazenadas em arquivo CVS para posterior consulta e análise.

A fim de reduzir o número de regras de associação obtidas no pós-processamento e melhor classificá-las foi proposto um modelo com o uso da ontologia de testes complementares das hepatites virais para generalizar os testes complementares fazendo com que o número de atributos seja reduzido no pré-processamento.

Um aplicativo construído em Java utilizando o framework Jena[19] e o *plugin* Pellet[20] foi construído para importar o modelo da ontologia e realizar a inferência dos pedidos.

A aplicação Infexmapp.jar foi executada passando como parâmetro o nome do arquivo OWL que será utilizado como referência para a pesquisa. No arquivo de configuração da aplicação foram identificadas as classes que representam o processo de diagnóstico das hepatites virais correspondendo à generalização dos exames complementares conforme axioma de equivalência modelado na ontologia.

A aplicação em questão instancia um modelo ontológico com suas classes, propriedades, relações com base na ontologia HVO.OWL salva em RDF. Esta aplicação lê os pedidos e exames complementares do banco de dados Postgree e cria as instâncias considerando a classe '*hvo.laboratory test'* e suas subclasses.

A identificação da subclasse a ser instanciada considerada a anotação *has_alternative_id* como forma de relacionar a representação LOINC com o teste complementar presente no pedido extraído da base de dado da empresa. A classe *hvo.laboratory_testing_encounter* é instanciada para cada pedido sendo associado as instâncias das subclasses de '*hvo.laboratory test'* através da propriedade de objeto '*is_associated'*. Caso algum exame complementar não possua uma classe correspondente este é desconsiderado sendo gravado na tabela de exames_não_inferidos.

[19]Disponível em:< https://jena.apache.org/>. Acesso em: 10.mai.2014.

[20]Disponível em:< https://github.com/complexible/pellet Acesso em: 10.mai.2014.

Com as instâncias importadas é executado o motor de inferência Pellet o qual identifica através do axioma de equivalência das classes de diagnóstico de hepatites virais, vide apêndice D, a ocorrência de cada tipo de hepatite que está sendo avaliada no pedido médico e retorna para a tabela de pedidos inferidos esta informação. Os exames complementares que não estão diretamente relacionados ao diagnóstico das hepatites virais são gravados na tabela de exame_nao_inferidos.

Após execução da aplicação, a relação de pedidos com a identificação de quais hepatites virais foi localizada é gravada na tabela de *pedido_inferência* e na tabela *exames_não_inferidos* o código dos exames complementares que não fazem parte do diagnóstico de hepatites virais.

QUADRO 4 – Relação de pedidos com generalização de testes complementares

Pedido	Exame 1	Exame2	Exame N	Exames generalizados
8000022	ELISAG	ELISAM	HTLV1	Exames Hepatite A
8000022	CBI	CHLA-M	FSH	Exames Hepatite C
8000058	HIV-EL	VD	CMM-ES	Exames Hepatite D
8000058	CMG-ES	CMM-ES	G	Exames Hepatite E
8000101	25-VD3	ELFC	CRE	Exames Hepatite G

Fonte: Elaborado pelo próprio autor.

Os dados obtidos foram importados de forma que fosse possível determinar regras por meio do algoritmo de mineração de dados por regras de associação Apriori.

Portanto, a base de Unidade e a Base de terceirizados foram reprocessadas considerando o resultado obtido com o motor de inferência Pellet considerando o axioma de equivalência sobre o diagnóstico de hepatites virais.

Quando na regra de associação obtida havia algum exame complementar não modelado na ontologia este foi incluído para que na fase de pós-processamento pudesse ser avaliada a similaridade semântica com os exames de hepatites virais.

Já na fase de pós-processamento, foi utilizado o cálculo de similaridade semântica entre o antecedente e o consequente para as regras de associação obtidas de forma a classificar as regras mais relevantes utilizando o modelo de Tversky (1977).

O algoritmo Apriori tradicional utiliza duas informações para a inferência de regras de associação: suporte e confiança. Contudo, nem sempre essas métricas são suficientes para determinar se padrões encontrados nos dados são significativos. Portanto, além de suporte e confiança foi utilizada a métrica *Lift* que terá seu valor recalculado com base na similaridade entre os exames complementares do antecedente e consequente para as regras de associação obtidas. A adoção do cálculo de similaridade entre o antecedente e o conseqüente indica o quanto mais próximo os testes complementares estão quando considerado as propriedades em comum, portanto, onde houver maior relação haverá maior relevância no padrão obtido.

Para a etapa da composição da base de conhecimento, foi necessário editar a saída do algoritmo Apriori. Os exames que compõe o antecedente foram separados para que pudesse ser calculada a similaridade semântica com o seu consequente. Portanto, com os resultados obtidos como saída do algoritmo Apriori, foram separados os antecedentes, consequente e o valor da métrica *Lift*. A Tabela 1 apresenta um exemplo da estrutura dos dados obtida.

TABELA 1 - Exemplo de regras de associação – Unidade Terceirizados

Regras	Suporte	Confiança	Lift
{exame31=ALUM,exame367=PTH} => {exame461=HEPATITE B}	0.014	1	1,10
{exame398=T4-RIE,exame417=TSH-B} => {exame461=HEPATITE B}	0.010	0.925	1,02
{exame162=ELISAG,exame273=HIV-ME} => {exame461=HEPATITE B}	0.012	0.954	1,05
{exame163=ELISAM,exame273=HIV-ME} => {exame461=HEPATITE B}	0.012	0.950	1,05

Fonte: Elaborado pelo próprio autor.

Ao *lift* da regra foi somado o valor obtido com o cálculo da similaridade entre o antecedente e consequente. A similaridade foi avaliada através de uma

aplicação Java que considera o modelo relacional de Tversky (Harispe, 2013), apresentado no item 2.2.3, conforme exemplificado abaixo:

Onde C e D são conceitos, ∏ é o operador de interseção de lógica descritiva, |C∏ D| é a cardinalidade da interseção dos conceitos, |C – D | e |D – C| resultam na cardinalidade da diferença do primeiro com o segundo.

Substituindo C e D por, respectivamente,

Exames complementares Diagnostico Hepatite vírus A ≡ Teste laboratorial ∏ ∃CausadoAgenteInfeccioso.(∃TemPresençaAgenteInfeccioso.Virus)

Exames complementares Diagnostico Toxoplasmose ≡ Teste Laboratorial ∏ ∃CausadoAgenteInfeccioso.(∃TemPresençaAgenteInfeccioso.Parasita)

Assim:

| Exames complementares Diagnostico Hepatite vírus A ∏ Exames complementares Diagnostico Toxoplasmose | = | Teste Laboratorial □ ∃CausadoAgenteInfeccioso | = 2

| Exames complementares Diagnostico Hepatite vírus A - Exames complementares Diagnostico Toxoplasmose | = | ∃TemPresençaAgenteInfeccioso.Virus | = 1

| Exames complementares Diagnostico Toxoplasmose - Exames complementares Diagnostico Hepatite vírus A | = |Parasita | = 1

Estes conjuntos formam as entradas para as funções utilizadas no modelo de Tversky (1).

$$s(a,b) = \frac{f(A \cap B)}{f(A \cap B) + \alpha * f(A - B) + \beta * f(B - A)} \qquad (1)$$

A parte dois analisa o contexto. Supondo que as características exclusivas do conceito C são mais importantes do que as do conceito D, com isso o fator α representando o contexto de C será 1, e o fator β, das características do conceito D será 0,5.

Com essas informações é possível avaliar a similaridade entre C e D com a função do Modelo Relacional (*Ratio Model)* de Tversky (1).

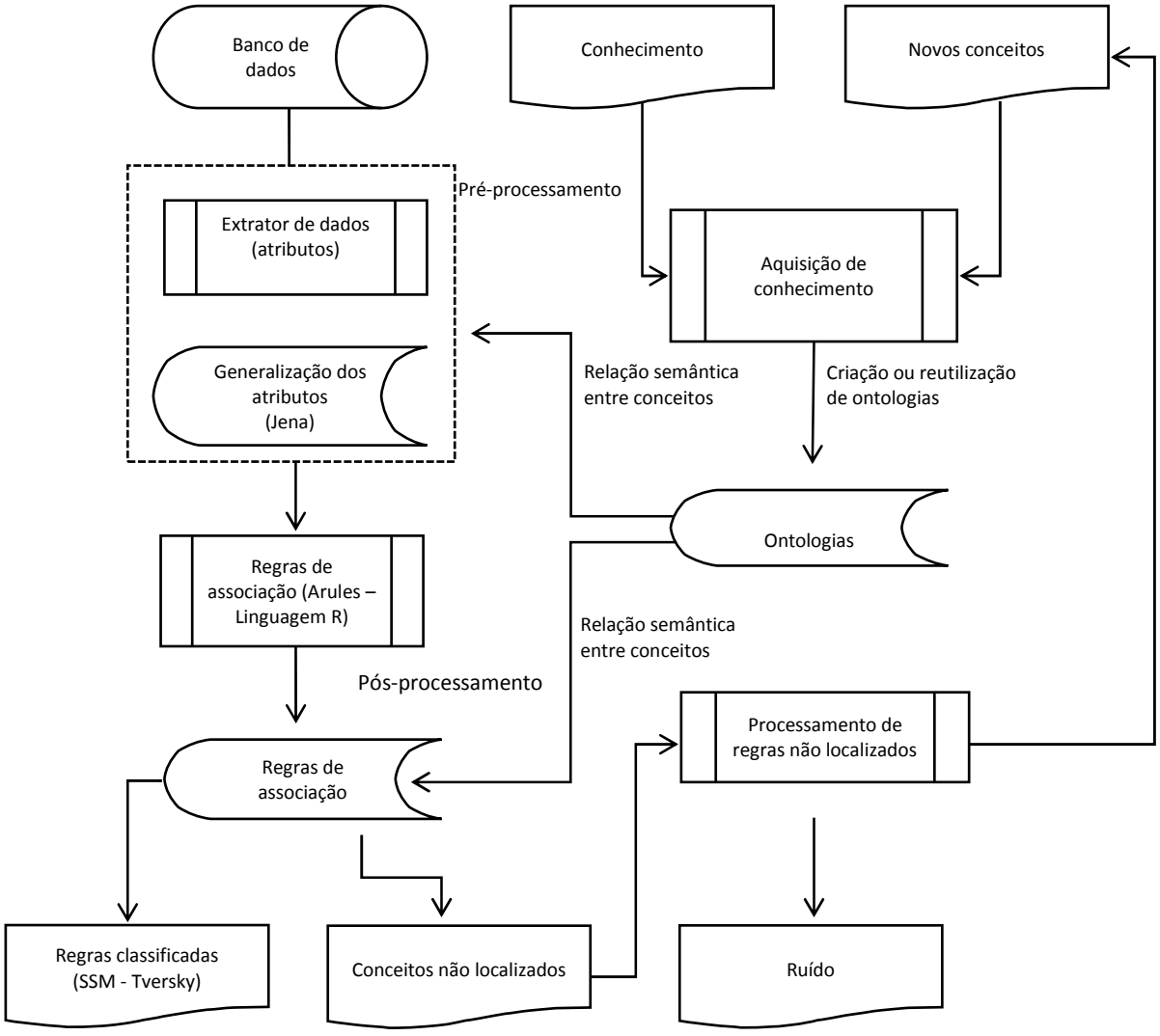
Com isso, a similaridade entre C e D é dada por:

S(C,D) = 2/(2 + 1*1 + 1*0,5) = 0,563.

As regras foram reclassificadas considerando o valor de *lift* readequado após soma do cálculo da similaridade semântica entre os conceitos.

A Figura 15 apresenta o modelo para generalização e classificação de regras de associação com o uso de ontologias na fase de pré e pós-processamento da mineração de dados.

FIGURA 15 – Modelo para extração de padrões com uso de ontologias



Fonte: Adaptado de Hamani, 2013.

3.3.6 Distribuição

Para validação foram selecionadas as 20 regras mais relevantes que foram apresentadas nas figuras 21 à 24 que representam o suporte e *lift* considerando as bases de unidades e terceirizados, mostrando as regras obtidas

sem o uso da ontologia e com o seu uso que foram validadas em entrevista conforme Apêndice B.

3.3.7 Softwares utilizados

Há algumas ferramentas de MD que auxiliam o processo de KDD. Neste trabalho, optou-se pelo uso do software Linguagem R devido licença livre (*open source*) e facilidade de uso em pesquisas acadêmicas nesta área.

Na etapa de MD foi utilizada uma amostra considerada representativa, sendo dividida em três fases:

Conjunto de Treinamento (*TrainingSet*): conjunto de registros usados no qual o modelo é desenvolvido, correspondendo a 10% dos registros;

Conjunto de Testes (*TestSet*): conjunto de registros usados para testar o modelo construído, correspondendo a 20% dos registros;

Conjunto de Validação (*ValidationSet*): conjunto de registros usados para validar o modelo construído, considerando os 70% dos registros restantes da amostra.

Os softwares utilizados para a criação e implementação da extração por regras de associação utilizando o algoritmo Apriori foram o RStudio v 0.98 e o pacote estatístico R (versão 3.1.3) e o sistema de gerenciamento de banco de dados PostgresSQL, versão 9.4.

Além das funções do pacote base do R, foram necessários outros pacotes como o pacote “Arules” (HASHER, 2005) para geração de regras de associação e na etapa de distribuição foi utilizado o pacote “arulesViz”[21] para geração dos gráficos apresentados para validação pelos especialistas.

Para generalização dos atributos na fase de pré-processamento foi utilizado o *framework* Jena e o racionador Pellet.

Jena é um *framework* em Java para construir aplicações para a web semântica que fornece classes e interfaces para criação e manipulação de RDF e ontologias baseadas em OWL. A API do Jena suporta alguns motores de inferência que são importados junto com o modelo da ontologia, arquivo OWL, e permite

[21] Disponível em:< https://cran.r-project.org/web/packages/arulesViz/index.html>. Acesso em: 1 .mar. 2014.

responder a consultas usando as inferências a partir dos axiomas existentes no OWL.

4 RESULTADOS E DISCUSSÃO

Um dos resultados desse projeto foi à construção da ontologia HVO que possibilitou a aquisição do conhecimento necessário à análise da base de dados dos testes laboratoriais obtendo regras de associação que agregassem conhecimento aos especialistas sobre o comportamento de prescrição médica. Além disso, a consolidação de atributos na fase de pré-mineração, ao generalizar os testes complementares relacionados à doença, permitiu reduzir o número de regras de associação obtidas.

O estudo propôs a utilização de ontologias, motores de inferência e o software Jena para realizar a poda e filtragem de dados (generalização) da relação de testes laboratoriais extraídos da base de dados da empresa de medicina diagnóstica objeto deste experimento. Ao analisar as relações que os termos compartilham é possível identificar quais testes laboratoriais estão relacionados à qual doença e fase considerando, portanto, a similaridade entre os termos como forma de generalizar os atributos na fase de pré-processamento.

Neste capitulo, são apresentados resultados da analise exploratória dos dados da base de conhecimento gerada na forma de regras de associação e, por fim, o resultado referente ao processo de validação das regras obtidas.

4.4.1 Modelagem ontológica

A elaboração da ontologia de exames complementares para diagnóstico de hepatites virais, HVO, se baseou nos sete passos da metodologia *Ontology Development* 101 (NOY; MCGUINNESS, 2001).

Passo 1: O escopo da ontologia é a modelagem dos testes laboratoriais de análises clínicas para diagnóstico de hepatites virais humanas tendo como referência para pesquisa as terminologias LOINC e SNOMED-CT, definidas, respectivamente, na portaria do Ministério da Saúde como padrões de interoperabilidade de exames laboratoriais e termos clínicos.

Passo 2: Ao analisar as ontologias biomédicas disponíveis no *OBO Foundry* e considerando os protocolos do ministério de saúde para diagnóstico de hepatites virais identificou-se a relevância em reutilizar a ontologia OGMS. A OGMS abordada por Scheuermann *et al.* (2009) descreve o quadro clínico sobre a

perspectiva da doença conforme Figura 10 mapeando o quadro terminológico que abrange as doenças, suas causas, manifestações, diagnóstico e outras entidades relacionadas. Com base nas pesquisas as bases LOINC e SNOMED-CT disponibilizada pelo Bioportal e tomando como referência a ontologia OGMS foi possível entender os termos, conceitos e as suas relações, bem como, o entendimento das classes que seriam reutilizadas para mapear o domínio de exames laboratoriais relacionados ao diagnóstico de hepatites virais. Além disso, fez-se necessário a pesquisa em fontes secundárias (literatura especializada, manuais de exames, protocolos do ministério da saúde) para relacionar as fases da doença, sinais, sintomas e a relação com os exames laboratoriais de acordo com o curso da doença que não ficaram claros quando considerado as bases LOINC e SNOMED-CT.

Através da ferramenta Ontofox[22] foi extraído trechos das ontologias DOID, FMA, IDO e OBI algumas classes relacionadas ao processo de diagnóstico das doenças que são relevantes para a construção da HVO.

Foi importado da ontologia DOID as classes e as suas subclasses que representar as doenças. Além disso, foi importado da ontologia IDO a classe "*Infectious Disease course*" e suas subclasses para representar o curso da doença infecciosa o qual influencia na prescrição de testes complementares. Na ontologia OBI foi importado as classes "*specimen*", "*analyte assay*" e "*organism*" e suas subclasses para representar, respectivamente, a espécime do material biológico, método e o agente causador da infecção. Para completar, foram importadas da FMA classes e subclasses que representam os órgãos e sistemas do corpo humano.

Passo 3: O levantamento dos termos relevantes para o desenvolvimento da ontologia foi validado pelas fontes primárias (os próprios especialistas) através do questionário do Apêndice A e relacionados na ontologia os testes laboratoriais disponíveis pela empresa de medicina diagnóstica.

Este levantamento foi validado através de entrevista com especialistas do laboratório onde o experimento está sendo realizado. Na entrevista foi possível validar a existência de exames complementares diretamente relacionados à doença e identificar outros exames inespecíficos. Os exames inespecíficos direcionam o diagnóstico médico podendo estar relacionado à avaliação da função do

[22] Disponível em:<http://ontofox.hegroup.org/>. Acesso em: 1 .mar. 2014.

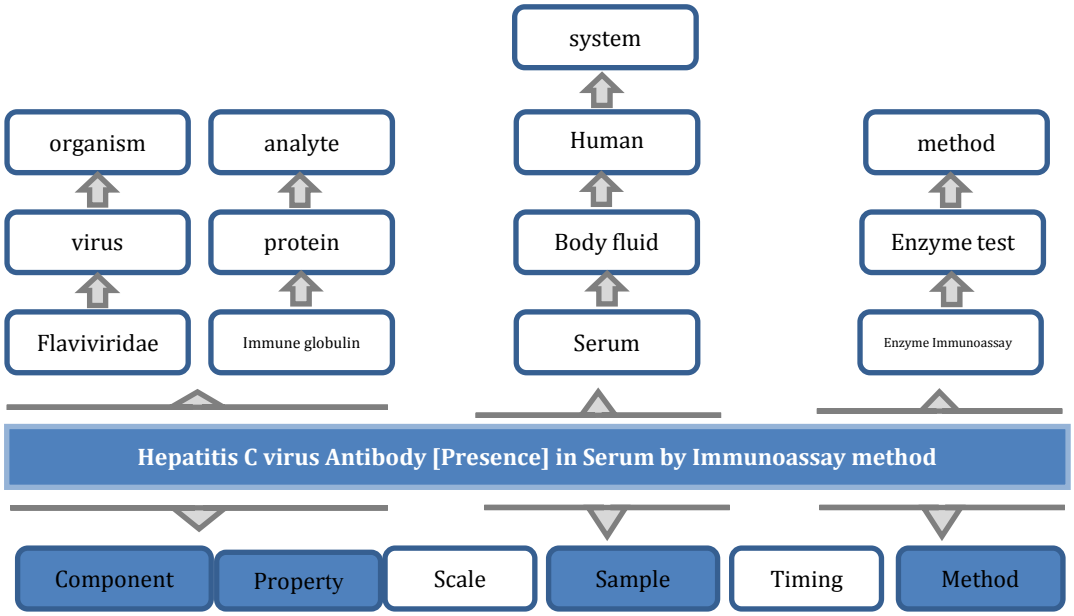
órgão/sistema onde a doença é mais presente ou para acompanhamento de outras funções do organismo.

Passo 4: Após validação com especialistas, considerando a ontologia OGMS com base para a modelagem, foram criadas as classes e hierarquia de classes da ontologia de testes complementares de hepatites virais humanas que foi nomeada HVO.

Passo 5: As propriedades das classes foram definidas considerando as relações entre os processos de atendimento laboratorial, diagnóstico da doença, testes laboratoriais, a doença e seu curso;

A OGMS foi adaptada para inclusão dos testes laboratoriais com base no método proposto por Eilbeck *et al.* (2013) que contempla a estrutura do LOINC considerando os atributos de identificação (ID LOINC, um nome, um gênero) e as quatro relações significativas para o escopo da pesquisa (um sistema, um analito[23], um método e um organismo). A Figura 16 apresenta um teste laboratorial com as suas propriedades distintivas na ontologia e seus correspondentes valores exclusivos em LOINC®.

FIGURA 16 - Exemplo classificação LOINC

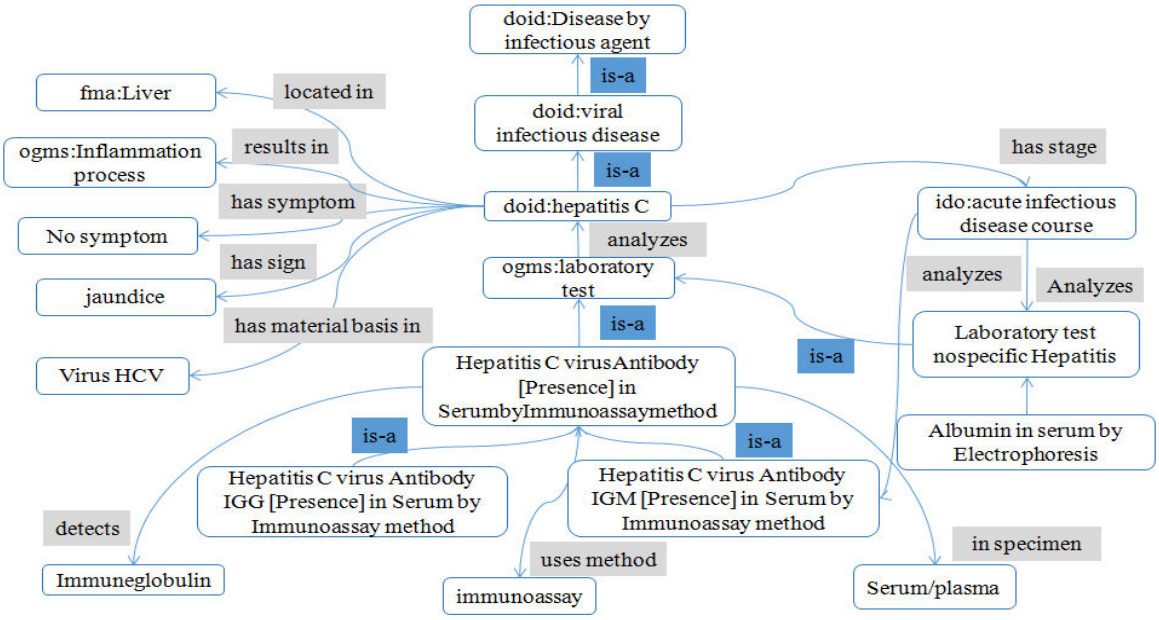


Fonte: Adaptado de EILBECK *et al.*, 2013.

[23] Analito– espécie química presente na amostra cuja concentração se deseja determinar em uma análise.

Além das relações especificadas no LOINC para definição dos exames laboratoriais, tomando como base o diagnóstico da doença proposto na OGMS, faz-se necessário a associação dos testes laboratoriais com a doença, seus sintomas e sinais, conforme apresentado na Figura 17, para identificar a relação entre os testes complementares e o diagnóstico médico.

FIGURA 17 - Exemplo classificação hepatite C



Fonte: Adaptado de EILBECK *et al.*, 2013.

Passo 6: As restrições das propriedades das classes foram criadas baseadas nos protocolos de diagnóstico das doenças relacionando o atendimento no laboratório e a relação de testes complementares para diagnóstico da doença.

Passo 7: Para validar o modelo ontológico foram criadas instâncias de alguns atendimentos laboratoriais considerando a lista de testes laboratoriais solicitados para diagnóstico de hepatites virais humanas.

Levando em consideração a relação de exames complementares relacionadas ao diagnóstico de hepatites virais, relacionadas na ferramenta RELMA/LOINC, e o conhecimento obtido com a aplicação do questionário do Apêndice A realizado com os especialistas. O desenvolvimento da ontologia é

iniciado pelo processo de extração de conhecimento, com a análise e relacionamento entre as terminologias LOINC e adotada pela empresa de medicina diagnóstica das tabelas de exames complementares, conforme trecho apresentado no Quadro 5.

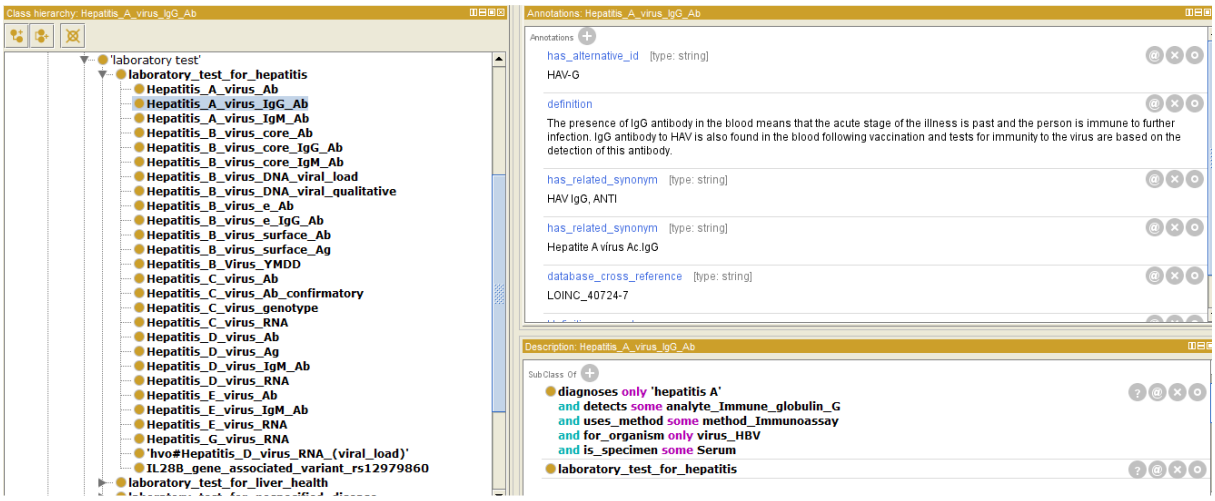
QUADRO 5 – De/ Para LOINC – Empresa medicina diagnóstica

Identificação	Nome exame	Método	Unidade	LOINC	Descrição Inglês	Descrição inglês resumida	Descrição português	Sistema	Método	Analitos
S/HAV-G	HAV IgG, ANTI	ELETROQUIMIO LUMINESCENCI A	-	40724-7	Hepatitis A virus IgG Ab [Presence] in Serum by Immunoassay	Hepatitis A virus IgG Ab	Hepatite A vírus Ac.IgG	Soro	EIA	Immune globulin G
S/HAV-M	HAV IgM, ANTI	ELETROQUIMIO LUMINESCENCI A	-	13950-1	Hepatitis A virus IgM Ab [Presence] in Serum or Plasma by Immunoassay	Hepatitis A virus IgM Ab	Hepatite A vírus Ac.IgM	Soro	EIA	Immune globulin M
S/HAVT	HAV TOTAL, ANTI	ELETROQUIMIO LUMINESCENCI A	-	13951-9	Hepatitis A virus Ab [Presence] in Serum by Immunoassay	Hepatitis A virus Ab	Hepatite A vírus Ac	Soro	EIA	Immune globulin
S/AU	HBsAg	ELETROQUIMIO LUMINESCENCI A	-	5196-1	Hepatitis B virus surface Ag [Presence] in Serum or Plasma by Immunoassay	Hepatitis B virus surface Ag	Hepatite B vírus superfície Ag	Soro	EIA	viral antigen
S/HBC-G	HBC IgG, ANTI	ELETROQUIMIO LUMINESCENCI A	-	40725-4	Hepatitis B virus core IgG Ab [Presence] in Serum by Immunoassay	Hepatitis B virus core IgG Ab	Hepatite B vírus core Ac IgG	Soro	EIA	Immune globulin G
S/HBCT	HBC TOTAL, ANTI	ELETROQUIMIO LUMINESCENCI A	-	13952-7	Hepatitis B virus core Ab [Presence] in Serum or Plasma by Immunoassay	Hepatitis B virus core Ab	Hepatite B vírus core Ac	Soro	EIA	Immune globulin
S/HBS	HBS, ANTI	ELETROQUIMIO LUMINESCENCI A	UI/L	10900-9	Hepatitis B virus surface Ab [Presence] in Serum by Immunoassay	Hepatitis B virus surface Ab	Hepatite B vírus superfície Ac	Soro	EIA	Immune globulin
S/HBC-M	HBC IgM, ANTI	ELETROQUIMIO LUMINESCENCI A	-	24113-3	Hepatitis B virus core IgM Ab [Presence] in Serum or Plasma by Immunoassay	Hepatitis B virus core IgM Ab	Hepatite B vírus core Ac IgM	Soro	EIA	Immune globulin M

Fonte: Elaborado pelo próprio autor.

Na presente tabela, constam alguns dos exames complementares relacionados ao diagnóstico de hepatites virais no LOINC e seu equivalente disponibilizado pela empresa de medicina diagnóstica objeto do estudo. Para cada exame complementar do LOINC foi criado uma subclasse associada à classe *OGMS: laboratory test*. Na propriedade de anotação das classes criadas foram relacionadas as informações referentes à identificação do exame complementar, informações adicionais provenientes da relação (*hvo: diagnoses*) à classe de exame complementar com a doença descrita na DOID (*doid:hepatitis A*) e às relações (*hvo:detects, hvo:uses_method, hvo: for_organism, hvo:is_specimen*) associadas , à classificação do exame no LOINC, conforme exemplificado na Figura 18.

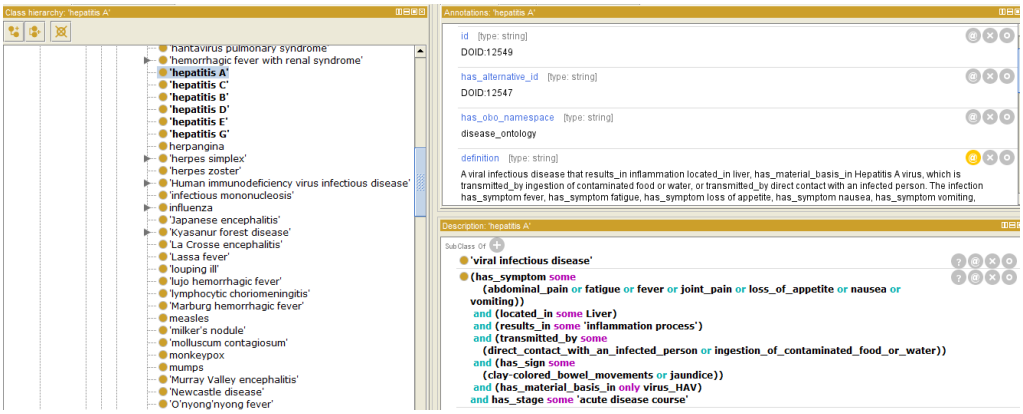
FIGURA 18 – Relação de exames complementares – Anotações e relações



Fonte: Elaborado pelo próprio autor.

Foram consideradas as classes descritas pela reutilização da ontologia DOID onde cada instância especifica as hepatites virais relacionando os sintomas (*doid: has_symptom*), sinais (*hvo:has_sign*), transmissão (*doid:transmitted_by*), agente causador (*doid: has_material_basis_in*) e com a FMA a relação referente ao órgão/sistema (*fma:located_in*) o qual a doença predominante ataca (*fma:liver disease e fma:liver*), conforme exemplificado na Figura 19.

FIGURA 19 – Hepatite A



Fonte: Elaborado pelo próprio autor.

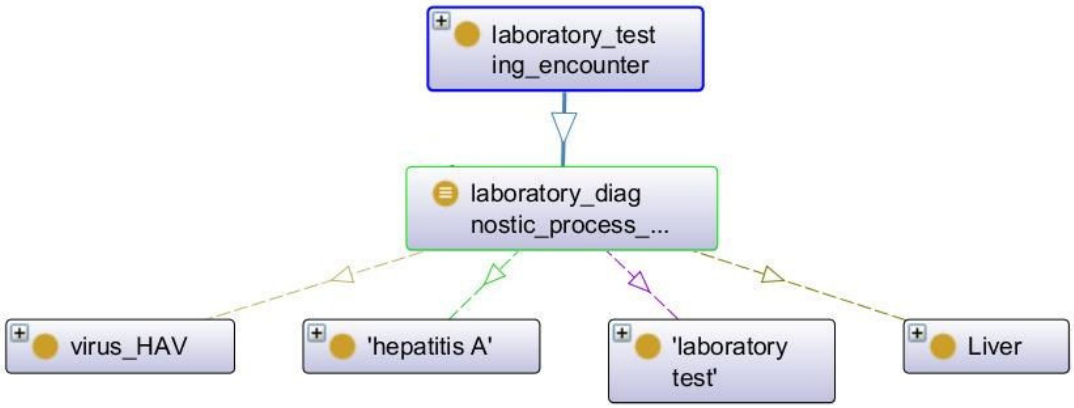
O atendimento ao paciente quando da apresentação da prescrição médica (*hvo:laboratory_test_encounter*) relaciona-se (*hvo: is_composed_of*) com os testes laboratoriais (*ogms:laboratory test*).

Neste axioma de equivalência (descrito por meio de restrições existenciais) é detalhada uma parte do processo de diagnóstico da doença sendo composta por, pelo menos, um pedido médico (nesse caso classes de *HVO: laboratory_testing_encounter*) que é composto de exames complementares *(ogms: laboratory test*) para diagnóstico da doença (*doid: hepatitis A*) e também pode haver algum teste laboratorial que avalia o estado de saúde (*HVO: diagnostic evaluation*) do fígado (*fma:liver*), conforme exemplificado na Figura 20. Assim, é assumido que pelo menos um exame complementar específico para avaliação de hepatite A deve estar presente no pedido médico para caracterizar a avaliação do diagnóstico da doença.

Exemplificando tal axioma:

Hvo:laboratory_diagnostic_process_ hepatitis_A equivalentto:
hvo:laboratory_testing_encounter
and (is_composed_of some
(‘laboratory test’
and (diagnoses only ‘hepatitis A’)))
and (is_composed_of min 0 (‘laboratory test’
and (diagnostic_evaluation some Liver)))

FIGURA 20 – Diagnóstico laboratorial para Hepatite A



Fonte: Elaborado pelo próprio autor.

4.4.2 KDD

Com base no conhecimento obtido com a elaboração da ontologia HVO foram extraídos da base de dados da empresa a relação de solicitação de exames que continham ao menos um dos exames relacionados diretamente com a prescrição para diagnóstico quando da hipótese de infecção por vírus das hepatites. Foram selecionadas as solicitações de testes laboratoriais dos meses de janeiro até março/2015 das unidades de atendimento e terceirização, que atendiam a regra descrita anteriormente, totalizando os valores apresentados na Tabela 2, para o conjunto de validação.

A amostra da Base Unidades possui 458 atributos, ou seja, 458 exames complementares distintos. Na amostra da Base de Terceirizados há 433 atributos, ou seja, 433 exames complementares distintos. Considerando a amostra das duas bases, há 465 exames complementares distintos.

Considerando a base de Unidades (suporte = 0,2) e base terceirizados (suporte =0,01) e confiança de 0,75 foram obtidos os resultados apresentados abaixo, sendo que a função para retirar as regras repetidas para a base de terceirizados apresentou erro devido tamanho da matriz a ser avaliada.

TABELA 2 - Atendimentos e regras de associação obtidas

Base	Atendimentos em que constam exames hepatites virais			Qtde regras obtidas	Qtde regras distintas
	Jan	Fev	Mar		
Unidades	927	819	10	2341	573
Terceirizados	7652	7698	17331	112673	221

Fonte: Elaborado pelo próprio autor.

Devido o alto consumo de memória da aplicação utilizando o framework Jena/Pellet foi necessário ajustar os parâmetros iniciais de alocação e dividir o processamento dos registros de forma que houvesse liberação de recurso.

O desenvolvimento da ontologia com os respectivos mecanismos de inferência para atuar na pré – mineração reduziu o número de atributos a serem minerados pelo algoritmo de regras de associação Apriori, à medida que, reduziu a quantidade de exames relacionados, ao diagnóstico direto do vírus das hepatites e exames inespecíficos para avaliação do curso das doenças. Considerando a mesma base de dados obteve-se a redução do número de atributos para 439 testes laboratoriais distintos. Por exemplo, para a base Unidades considerando o suporte de 0,2 e confiança de 0,75 foram obtidas 881, retirando-se as regras redundantes restaram 258 regras, conforme Tabela 3.

TABELA 3 - Atendimentos e regras de associação obtidas com ontologia

Base	Atendimentos em que constam exames hepatites virais			Qtde regras obtidas	Qtde regras distintas
	Jan	Fev	Mar		
Unidades	927	819	10	881	258
Terceirizados	7652	7698	17331	3672	115

Fonte: Elaborada pelo próprio autor.

Ao criar *itemsets* genéricos referenciando a doença em que ocorrem é possível reduzir os atributos na fase de pré-mineração melhorando a desempenho do minerador, devido redução das combinações e, consequentemente, do número

de resultados no pós-processamento facilitando a análise dos resultados e classificação.

O formato de uma Regra de Associação pode ser representado como uma implicação LHS -> RHS, em que LHS (antecedente) e RHS (consequente) são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens. Com saídas filtradas (RHS - consequente da regra) por uma das classes correspondente as hepatites virais (hepatites A,B,C,D,E e G) o número de regras de associação é reduzido para 75 e 129, respectivamente, unidades e terceirizados, conforme Tabela 4.

TABELA 4 - Atendimentos e regras de associação obtidas com ontologia e filtro por RHS (hepatites A,B,C,D,E e G)

Base	Atendimentos que constam exames hepatites virais			Qtde regras obtidas	Qtde regras distintas
	Jan	Fev	Mar		
Unidades	927	819	10	181	75
Terceirizados	7652	7698	17331	435	129

Fonte: Elaborada pelo próprio autor.

Com isso foi possível a redução das regras resultantes da mineração de dados, propiciada pela redução das possibilidades combinatórias dos atributos. Uma segunda contribuição é que a generalização dos termos possibilita um maior significado aos resultados capaz de orientar o processo de análise na fase de pós-mineração. Nesse aspecto, este método exerce uma função de métrica subjetiva dos resultados obtidos. Com isso se reduz o número de regras de interesse permitindo focar nas regras de relevância que não são conhecidas a princípio. Esta métrica associada com as técnicas objetivas de e suporte, confiança e *lift* utilizadas pelo algoritmo Apriori permitiram a construção de outros critérios de poda e classificação dos resultados na fase de pós-mineração com o aprimoramento da ontologia existente.

Com as regras geradas foi possível manter apenas as regras que apresentavam padrões com ocorrência de exames para diagnostico das hepatites virais A, B, C, D, E ou G, e posteriormente, classificar as regras mais relevantes considerando as relações entre as hepatites virais, conforme exemplificado na Tabela 5. A relação da prescrição de testes laboratoriais considerando a doença possibilitou relacioná-los e assim, generalizar e criar apenas um atributo que

representará os testes laboratoriais de cada hepatite viral. Cabe ressaltar que é comum a infecção por mais de uma hepatite viral e que os sintomas/sinais são iguais o que torna normal a ocorrência de testes laboratoriais para estas doenças na mesma solicitação médica permitindo uma maior redução no número de testes laboratoriais (atributos utilizados na mineração) em função desta relação.

TABELA 5 – Testes laboratoriais generalizados

Regras	Suporte	Confiança	Lift
{CA,FE,FERRI} => { HEPATITE B}	0.014	1	3.012
{CA,FE,IST} => { HEPATITE B}	0.012	1	3.012
{CA,FE,NA} => { HEPATITE B}	0.012	1	3.012
{CA,FE,K} => { HEPATITE B}	0.012	1	3.012

Fonte: Elaborada pelo próprio autor.

A partir do conjunto final de dados, já tratados, foram determinadas as regras de associação mais relevantes, conforme exemplificado na tabela 6, através da classificação por similaridade semântica entre antecendente e consequente, considerando a semelhança entre as doenças.

TABELA 6 - Regras de associação classificadas - Lift + SSM

Antecedente	Consequente	Lift	SSM	Resultado
Glicose, HIV 1 e 2 Ab Ag, TSH-B	Hepatitis C	1,24	0,238	1,478
Glicose, Urina rotina, VDRL	Hepatitis B	1,04	0,285	1,325

Fonte: Elaborado pelo próprio autor.

4.4.3 DISCUSSÃO

As técnicas de mineração de dados por regras de associação com o uso de ontologias foram adotadas para identificar padrões mais relevantes relacionados à prescrição médica de exames complementares. O uso da ontologia de domínio no processo de KDD possibilitou reduzir o número de regras e aumentar a precisão e relevância dos padrões obtidos.

As ontologias permitiram descrever a estrutura complexa do domínio biomédico possibilitando que as dificuldades oriundas do aumento na quantidade de dados em nível exponencial e da ausência de ferramentas para descoberta de conhecimento em domínios especializados sejam minimizadas. O uso de ontologias possibilitou automatizar o cálculo das relações entre os termos..

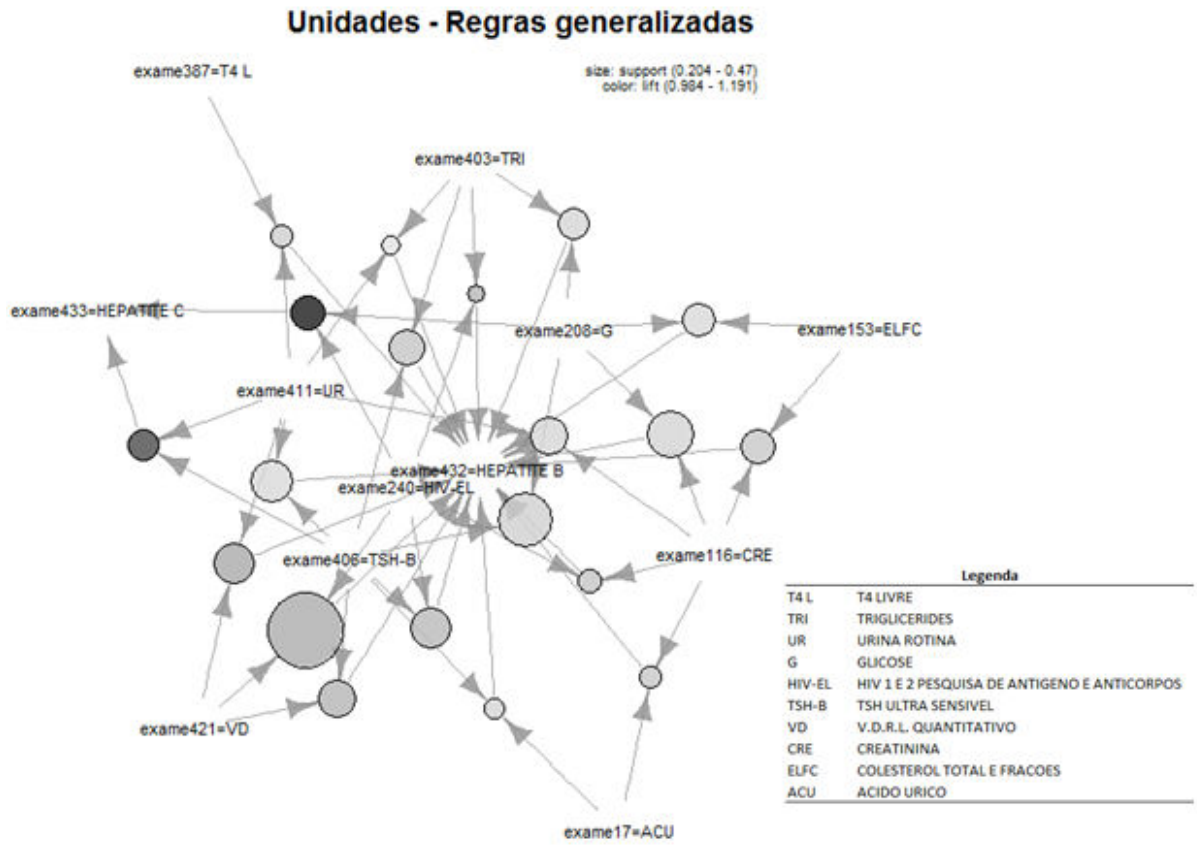
A utilização de ontologias nas fases de pré e pós- processamento à MD por regras de associação agregou o conhecimento de domínio e propiciou suporte semântico auxiliando o analista a explicar e interpretar as regras obtidas; possibilitando categorizar e classificar as regras geradas através da generalização e poda.

Com a generalização foi possível descartar os testes complementares que estavam diretamente associadas às hepatites virais e assim, dar mais importância aos outros que estão relacionados às demais doenças ou são exames de rotina.

Já o cálculo de similaridade semântica permite classificar os padrões com maior relevância descartando, por exemplo, testes complementares de rotina.

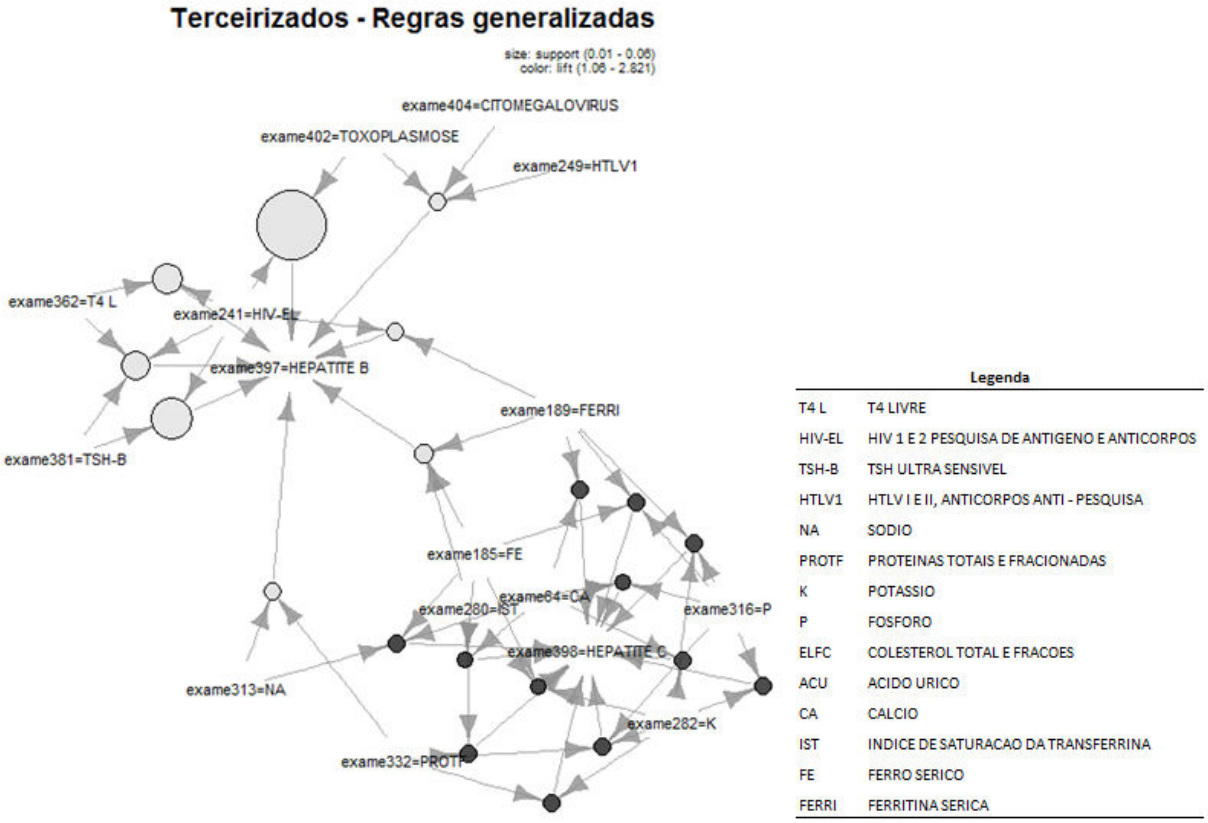
Considerando as regras extraídas após generalização dos exames de hepatites virais foram representadas as 20 mais relevantes, com base no *lift*, nas Figuras 21 e 22. O suporte é representado pela dimensão da bola e o *lift* pela cor.

FIGURA 21 – Unidades – Regras generalizadas



Fonte: Elaborada pelo próprio autor.

FIGURA 22 – Terceirizados – Regras generalizadas

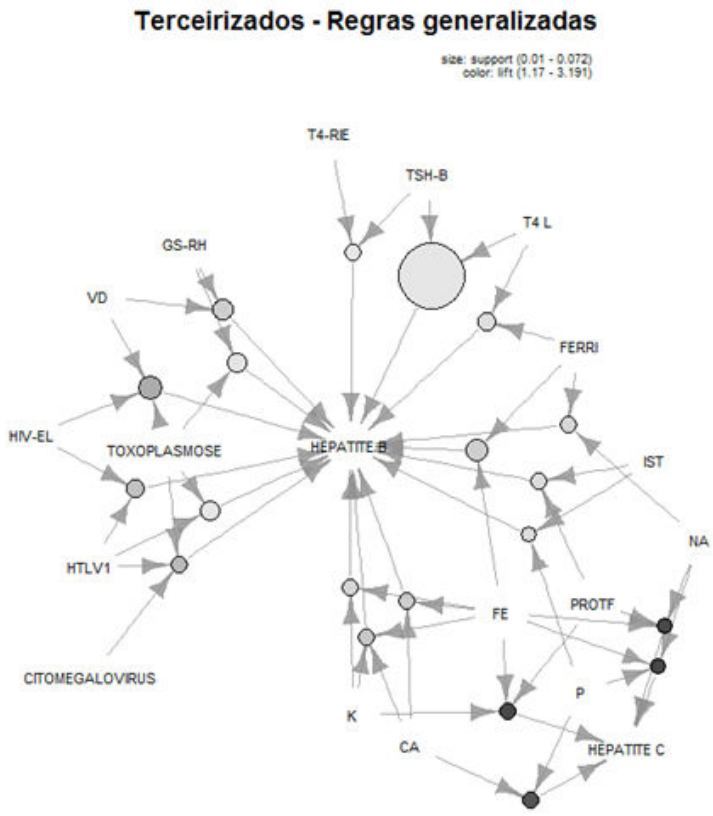


Fonte: Elaborada pelo próprio autor.

Considerando a similaridade através do modelo relacional de Tversky, com o contexto dado pela descrição anterior, foram obtidos novos valores de *lift* calculados a partir da similaridade entre os atributos do antecedente e o consequente, no caso, as hepatites virais.

Para cálculo da similaridade semântica entre os conceitos, os testes laboratoriais, foram consideradas as características relacionadas à doença, seus sintomas, tipo de agente causador, local ou sistema do corpo humano infectados, sendo os resultados obtidos após o cálculo representados pelas Figuras 23 e 24.

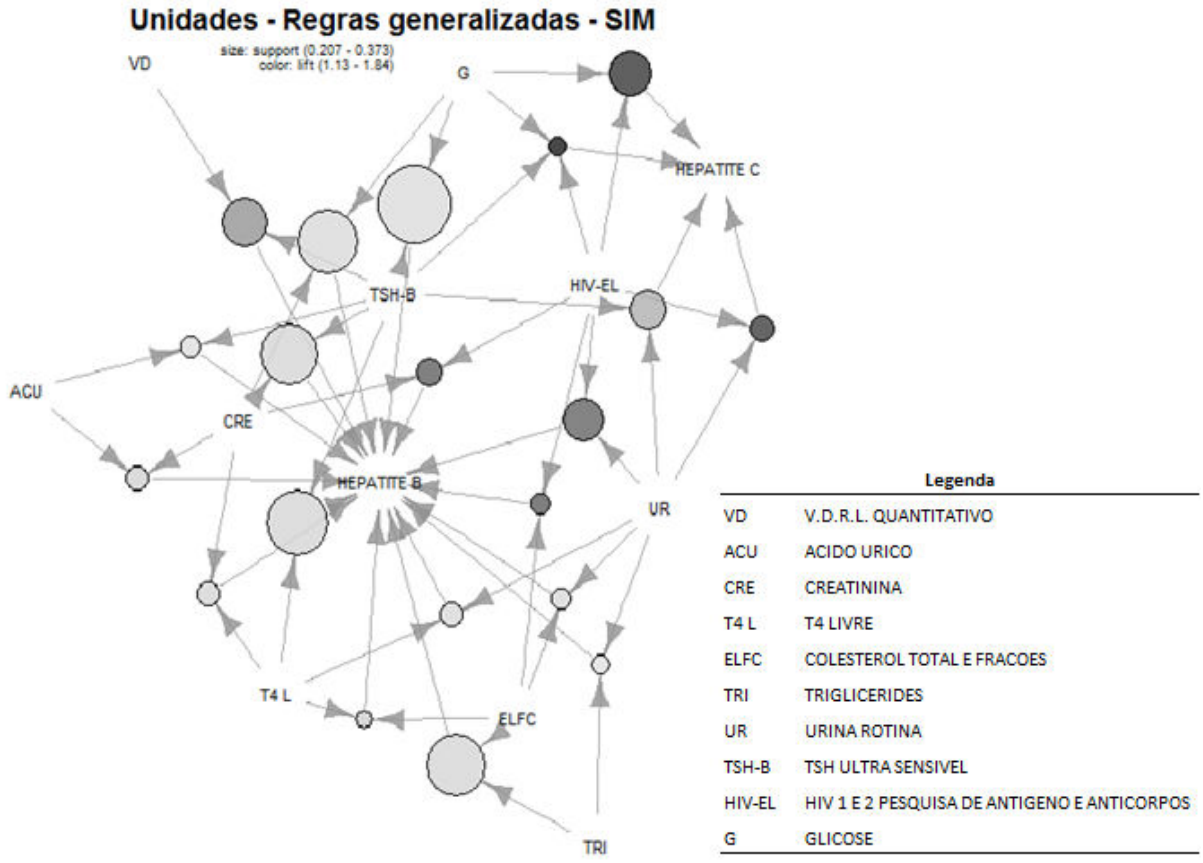
FIGURA 23 – Terceirizados – Regras generalizadas e com lift recalculado



Legenda	
GS-RH	GRUPO SANGUINEO + FATOR RH/DU
VD	V.D.R.L. QUANTITATIVO
HIV-EL	HIV 1 E 2 PESQUISA DE ANTIGENO E ANTICORPOS
HTLV1	HTLV I E II, ANTICORPOS ANTI - PESQUISA
T4 L	T4 LIVRE
TSH-B	TSH ULTRA SENSIVEL
NA	SODIO
PROTF	PROTEINAS TOTAIS E FRACIONADAS
K	POTASSIO
P	FOSFORO
ELFC	COLESTEROL TOTAL E FRACOES
ACU	ACIDO URICO
CA	CALCIO
IST	INDICE DE SATURACAO DA TRANSFERRINA
FE	FERRO SERICO
FERRI	FERRITINA SERICA

Fonte: Elaborada pelo próprio autor.

FIGURA 24 – Unidades – Regras generalizadas e com lift recalculado



Fonte: Elaborada pelo próprio autor.

Para validação dos resultados obtidos foram realizadas entrevistas com os especialistas do laboratório (biomédicos e farmacêuticos) onde os gráficos acima foram mostrados, apêndice B, e questionado se os resultados obtidos continham informações relevantes sobre a prescrição de testes complementares quando considerando o universo do diagnóstico de hepatites virais.

Com base nas entrevistas foram pontuadas as seguintes considerações:

- Com a readequação dos valores de *lift* foi possível classificar as regras interessantes priorizando a relação com a doença;
- Os padrões extraídos confirmam as colocações dos especialistas sobre a existência de prescrição de exames classificados como rotina, conforme prática médica, principalmente na base de atendimento das unidades;
- Existência de testes laboratoriais normalmente prescritos em medicina ocupacional na base de terceirização, principalmente, relacionados à hepatite C;

- Existência de testes laboratoriais relacionados à outras doenças infecciosas na base de terceirização, principalmente, relacionados à hepatite B;

Com os resultados obtidos foi possível observar padrões de prescrição médica que não estão diretamente atreladas às hepatites virais, mas que indicam um comportamento do prescritor em realizar um diagnóstico diferencial de acordo a situação e outras possíveis doenças relacionadas com base na anamnese.

Com a entrevista foi possível identificar oportunidades para avaliação de padrões conforme idade, região e sexo para identificar se as condições observadas podem ter comportamento distinto de acordo com estes atributos.

É importante ressaltar que modelar um domínio especializado como a área biomédica de forma mais ágil e precisa necessita da presença de especialistas. Portanto, investir na elaboração de outras ontologias biomédicas é importante para condução de mais trabalhos de recuperação da informação com uso do conhecimento prévio do domínio.

É importante ressaltar que durante modelagem na ferramenta Protegé foram avaliadas pacotes (*plugin*) para extração de regras como a DLQuery e SPARQL, mas devido a maior complexidade e pouca documentação optou-se por utilizar os mecanismos de motor de inferência disponíveis.

No entanto, a utilização do framework Jena com o motor de inferência Pellet também apresentou alguns problemas em função da documentação limitada e poucos exemplos do seu uso o que exigiu um tempo maior para o aprendizado. Durante a execução do motor de inferência para generalização dos testes complementares foi necessário dividir a execução dos registros devido problemas de alocação de memória.

Optou-se pela utilização da linguagem R para processamento das regras de associação pelo algoritmo Apriori devido boa documentação, possibilidade de uso de vários pacotes e grande uso pela comunidade KDD.

Portanto, esta pesquisa utilizou alguns documentos, ontologias e entrevistas para validar o conhecimento do domínio mas tem ciência que para aprimorar os resultados precisa de mais aprofundamento no processo de diagnóstico médico e dos testes complementares disponíveis.

Escrever sobre um experimento que envolve ciência da informação, ciência da computação e medicina diagnóstica exigiu um grande esforço de

pesquisa. Este trabalho, busca mostrar uma nova forma de pensar a resolução de problemas de recuperação da informação sobre uma perspectiva diferente.

5 CONCLUSÕES E TRABALHOS FUTUROS

Identificar o mercado e se posicionar corretamente na oferta de produtos e serviços constituem um grande desafio para os tomadores de decisão. A partir do entendimento das necessidades dos clientes é possível promover ações para direcionar as estratégias de venda e marketing.

Com os resultados obtidos demonstrou-se, na prática, como as diversas tecnologias ligadas ao processo de KDD e organização do conhecimento podem apoiar as tomadas de decisões, de forma a entender o comportamento médico quando da solicitação de testes complementares ao diagnóstico.

De acordo com os experimentos, foi possível obter resultados satisfatórios para responder ao objetivo do trabalho que é aprimorar o processo de recuperação de informação com a obtenção de regras de associação mais relevantes para os usuários com o uso de ontologias.

A ontologia de aplicação desenvolvida representou os testes LOINC® para hepatites virais em alto nível por entender que neste experimento esta classificação é suficiente na avaliação da relação das regras de associação e não pretende esgotar todos os termos e conceitos relacionados ao diagnóstico da doença e dos exames laboratoriais relacionados. Estender a OGMS aproxima os testes laboratoriais ao ciclo de diagnóstico da doença e, consequentemente, permite identificar as correlações entre os exames. Esta abordagem é, por conseguinte, extensível aos outros exames que forem identificados no processo de mineração por regras de associação e que inicialmente não estão relacionados de forma primária ao diagnóstico de hepatites cabendo avaliação e readaptação da ontologia. É importante salientar a importância de estender estes modelos para outras doenças infecciosas, pois a análise das regras de associação demonstrou forte associação com outros testes laboratoriais para doenças ocasionadas por vírus, por exemplo, HIV, rubéola e citomegalovirus.

Ao descrever cada teste utilizando princípios diferenciadores foi possível incluir os testes dentro de uma hierarquia de geral ao específico, proporcionando assim uma fonte de automatização da consulta para poda de regras e também para análise de relevância ao contexto.

É importante destacar dois pontos de atenção sobre o uso de similaridade semântica com o uso de ontologias: primeiramente, a relevância de padrões extraídos por técnicas de identificação de similaridade semântica entre termos é altamente dependente da construção e validação das ontologias, portanto, é importante que as ontologias de domínio utilizadas tenham sido validadas por especialistas do negócio; segundo, a abordagem no processo de KDD precisa de maior aprofundamento sobre os algoritmos não se restringindo apenas as relações ‘é um’ e ‘parte de’ melhorando uso da semântica formal proposta pelas ontologias na aplicação nos métodos computacionais da mineração de dados.

Com os resultados obtidos é possível mostrar como as diversas tecnologias ligadas ao processo de KDD e organização do conhecimento pode apoiar a tomada de decisão. A inclusão de embasamento semântico (ontologias) permite que informações estratégicas sejam descobertas e transformadas em conhecimento relevante para compreensão de forma a entender, por exemplo, o comportamento médico quando da solicitação de testes complementares ao diagnóstico médico.

A escolha de algoritmos adequados para cálculo da similaridade semântica de base ontológica entre os termos para obtenção de padrões sem redundância aprimorando a revocação e precisão tornou-se um grande desafio no processo de recuperação da informação em base de dados.

Algumas perspectivas podem ser abordadas em trabalhos futuros. Em primeiro lugar, foco no enriquecimento da ontologia com novos conceitos e equivalência entre os testes complementares e as doenças para maior generalização dos atributos. Conforme mencionado pelos especialistas alguns testes laboratoriais presentes nas regras são considerados de rotina não estando relacionados diretamente a uma doença especifica, portanto, precisa ser avaliada sua relevância quando da generalização dos atributos. Apesar da similaridade semântica para estes testes de rotina ser pequena, se comparados às hepatites virais, devido alto suporte e confiança possuem um valor inicial de *lift* que fez com que permanecessem na lista de regras relevantes.

Em segundo lugar, avaliar e desenvolver um algoritmo de similaridade semântica para a etapa de filtragem das regras na fase de classificação das regras que considere as especificidades do contexto em que este trabalho se insere.

Finalmente, investigar ontologias existentes que relacione a doença e seus testes complementares de forma a automatizar o mapeamento entre entes conceitos aprimorando a generalização dos atributos, visto que este universo é muito abrangente e exige conhecimento especializado. Com o aumento da procura por métodos computacionais para resolver o problema da extração de padrões em grandes massas de dados, *Big Data*, o uso de ontologias pode facilitar a organização e recuperação das informações ajudando a resolver os problemas de volume, variedade e veracidade.

REFERÊNCIAS

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. N. Mining association rules between sets of items in large databases. In: 20th VLDB CONFERENCE, 1993, Santiago. **Proceedings…**Ed. P. Buneman e S. Jajodia, Santiago: ACM Press Santiago, 1993. p. 207–216. Disponível em: <http://rakesh.agrawal-family.com/papers/vldb94apriori.pdf.> Acesso em: 1 mar. 2014.

ALMEIDA, M. B.; BAX, M. P. Uma visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Infomação,** Brasília, v. 32, n. 3, p. 7-20, 2003.

ALMEIDA, M. B. *et al.* Aportes linguísticos no apoio a construção de ontologias*.* **Liinc em Revista***,* v.6, n.2, p.384-410, set., 2010.

ALMEIDA, M. B. Revisiting ontologies: a necessary clarification. **Journal of the American Society of Information Science and Technology**. v. 64, n. 8. p. 1682-1693, 2013.

ALMEIDA, M.; SOUZA, R.; FONSECA, F. Semantics in the semantic web: a critical evaluation. **Knowledge Organization**, v. 38, n. 3, p. 187–203, 2011.

THE APACHE SOFTWARE FOUNDATION. Apache Jena. Disponível em:< https://jena.apache.org/>. Acesso em: 1 mar. 2014.

BAADER, F. *et al*. (Eds.). **The description logic handbook**: theory, implementation, and applications. Cambrigde: Cambridge University Press, 2003.

BAADER, F.; NUTT, W. **Basic description logics**. 2003. Disponível em: <https://www.inf.unibz.it/~franconi/dl/course/dlhb/dlhb-02.pdf>. Acesso em: 1 fev. 2015.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval.** New York: ACM Press, 1999.

BASTOS, A. **Uma abordagem ontológica baseada em informações de contexto para representação de conhecimento de monitoramento de sinais vitais humanos.** 2013.129 f. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal de Goiás, Goiânia, 2013. Disponível em: <http://www.inf.ufg.br/mestrado/sites/www.inf.ufg.br.mestrado/files/uploads/Dissertacao_Alexsandro.pdf>. Acesso em: 1 mar. 2014.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v.284, n.5, p. 28-37,2001.

BREITMAN, K**. Web semântica:** a internet do futuro. Rio de Janeiro: LTC, 2005.

BRASIL. Ministério da Saúde. *Biblioteca Virtual em Saúde*. Disponível em: <http://bvsms.saude.gov.br/bvs/saudelegis/gm/2011/prt2073_31_08_2011.html>. Acesso em: 1 mar. 2014.

CAMILO, C.O.; SILVA, J.C. 2010. **Um estudo sobre a interação entre mineração de dados e ontologias**. Goiás: [S.l.], 16 p. Relatório técnico. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_002-09.pdf>. Acesso em: 1 out. 2013.

CAPILHEIRA, M. F.; SANTOS, I. S. Epidemiologia da solicitação de exame complementar em consultas médicas. **Revista de Saúde Pública,** v. 40, n. 2, p. 289-297, 2006.

CAO, Longbing. Domain-driven data mining: challenges and prospects. **Transaction on Knowledge And Data Engineering**, v. 22, n. 6, june, p. 755-769, 2010.

CAO, Longbing. **Introduction to domain driven data mining**. Disponível em: <http://staff.it.uts.edu.au/~lbcao/publication/dmba-dddm.pdf>. Acesso em: 1 mar. 2014.

CHAPMAN, P. **CRISP-DM 1.0 step-by-step data mining guide**. United States of America, 2000. Disponível em: <http://www.crisp-dm.org>. Acesso em: 1 mar. 2014.

COELHO, E.M.P. **Ontologias difusas no suporte à mineração de dados**: aplicações na Secretaria de Finanças da Prefeitura Municipal de Belo Horizonte Belo Horizonte**,** 2012. Tese (Doutorado em Ciência da Informação)– Departamento de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2012.

COELHO, K. C.; ALMEIDA, M. B. Aquisição de conhecimento para construção de ontologias: uma proposta de roteiro metodológico aplicado ao contexto da hematologia. **Encontros Bibli**: revista eletrônica de biblioteconomia e ciência da informação, Santa Catarina, v. 17, n. 35, p. 47-74, 2012.

COUTO,F.; SILVA, M. Disjunctive shared information between ontology concepts: application to Gene Ontology. **Journal of Biomedical Semantics**, v. 2, n. 5, 2011. Disponível em: < http://dx.doi.org/10.1186/2041-1480-2-5> . Acesso em: 1 mai. 2014.

DALFOVO, O.; Amorim, S. N. **Quem tem informação é mais competitivo**: o uso da informação pelos administradores e empreendedores que obtém vantagem competitiva. Blumenau: Acadêmica, 2000.

DISEASE ONTOLOGY. Disponível em: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=disease_ontology> . Acesso em: 1 mar. 2014.

EILBECK, K. ; JACOBS, J. ; STAES, C.J., **Exploring the use of ontologies and automated reasoning to manage selection of reportable condition lab tests from LOINC**, 2013. Disponível em: < http://ieeexplore.ieee.org/xpl/abstractAuthors.jsp?arnumber=6480135> . Acesso em: 1 maio. 2014.

FAYYAD, U. M. *et al*. **Advances in knowledge discovery and data mining**. [S.l]: Springer-Verlag Berlin Heidelberg, 1996.

FERRAZ, I. **Ontology in association rules**.[S.l.: s.n.] Disponível em: < http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3786067/>. Acesso em: 1 mai. 2014.

FOUNDATIONAL model of anatomy. Disponível em: <http://sig.biostr.washington.edu/projects/fm/>. Acesso em: 1 fev.2015.

GAN, M.; DOU, X.; JIANG, R. From ontology to semantic similarity: calculation of ontology-based semantic similarity. **The Scientific World Journal**, 2013. Disponível em: <http://www.hindawi.com/journals/tswj/2013/793091/>. Acesso em: 10 mai. 2014.

GELAIM, T. A. **Similaridade em lógica descritiva**. Santa Catarina: Universidade Fedarl de Santa Catarina, 2013. Disponível em: < https://projetos.inf.ufsc.br/arquivos_projetos/projeto_1474/MonografiaThiagoAngeloGelaim.pdf>. Acesso em: 1 fev. 2015.

GIL, A.C. **Métodos e técnicas de pesquisa social.** 4 ed. São Paulo: Atlas, 1994. 207 p.

GRENON, P.; SMITH, B. SNAP and SPAN: towards dynamic spatial. **Spatial Cognition & Computation**, v.4, n.1, p. 69-104, 2004. Disponível em: <http://ontology.buffalo.edu/smith/articles/SNAP_SPAN.pdf>. Acesso em: 12 abr. 2012.

GRUBER, T.R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. Revision. In: INTERNATIONAL WORKSHOP ON FORMAL ONTOLOGY, 1993, Stanford. **Proceedings…** Stanford: Stanford Knowledge Systems Laboratory, 1993. Disponível em: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.3207&rep=rep1&type=pdf>. Acesso em: 1 out. 2013.

GRUBER, T. **What is an ontology?**.1993. Disponível em: <http://wwwksl.stanford.edu/kst/what- is-an-ontology.html>. Acesso em: 1 out. 2013.

GUARINO, N. Formal ontology and information systems. In:_____**Formal ontology in information systems**. Amsterdam: IOS Press, 1998. p. 3-15.

GUARINO, N.; GIARETTA, P. **Ontologies and knowledge bases**: towards a terminological clarification. [S.l.: s.n.], [ca. 2000]. Disponível em:

<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.320.8006&rep=rep1&type=pdf>. Acesso em: 1 out. 2013.

GONÇALVES, E.C. Regras de associação e suas medidas de interesse objetivas e subjetivas. **INFOCOMPJournal of Computer Science**, 2005. Disponível em:<http://www.dcc.ufla.br/infocomp/artigos/v4.1/art04.pdf>. Acesso em: 1 mar. 2014.

GONÇALVES, E.C. Data mining com a ferramenta WEKA. In: III FÓRUM DE SOFTWARE LIVRE DE DUQUE DE CAXIAS, 2011, Duque de Caxias. **Anais...** Duque de Caxias: Escola Nacional de Ciências Estatísticas IBGE/ENCE. 2011. Disponível em: <http://forumsoftwarelivre.com.br/2011/arquivos/palestras/DataMining__Weka.pdf>. Acesso em: 1 out. 2013.

GONÇALVES, L. P. F. Um estudo sobre a confiabilidade de ferramentas de mineração de dados. **Revista do CCEI - Centro de Ciências da Economia e Informática**, v.8, n.14, p. 37-47, ago. 2004.

HAN, J.; KAMBER, M. **Datamining**: concepts and techniques. 2006. Disponível em: <http://web.engr.illinois.edu/~hanj/bk2/slidesindex.ht>. Acesso em: 1 mar. 2014.

HAMANI, M. *et al.* Unexpected rules using a conceptual distance based on fuzzy ontology. **Journal of King Saud University - Computer and Information Sciences**, v. 6, n. 1, p. 99-109, 2014. Disponível em: < http://www.sciencedirect.com/science/article/pii/S1319157813000141> Acesso em: 1 mar. 2014.

HARISPE, Sébastien. A framework for unifying ontology-based semantic similarity measures: a study in the biomedical domain. **Journal of biomedical informatics**, v. 48, p. 38-53, 2013. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1532046413001834>. Acesso em: 1 fev. 2015.

KASAMA, D.; ZAVAGLIA, C.; BARCELLOS, G. Do termo à estruturação semântica: representação ontológica do domínio da Nanociência e Nanotecnologia utilizando a Estrutura Quali. **Linguamática**, v. 2, n. 3, p. 43–58. 2010. Disponível em: < http://www.jourlib.org/search?kw=%20estrutura%20Qualia%20&searchField=keyword >. Acesso em: 1 mai. 2014.

HASHER,M.; HORNIK, K.; GRUN, B.; BUCHTA, C. **Introduction to arules**: computational environment for mining association rules and frequent item sets. 2007. Disponível em: <http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>. Acesso em: 10 jan. 2015.

LOGICAL Observation Identifiers Names and Codes. Disponível em: <https://loinc.org/>. Acesso em: 1 mar. 2014.

LUZ, Daniel. **Implementação de uma ferramenta de data mining para o auxílio à tomada de decisão**: caso de uma cadeia de suprimentos. 2003. 48 p. Monografia. (Curso de Engenharia de controle e automação industrial) - Universidade Federal de Santa Catarina, Santa Catarina, 2003. Disponível em: <http://www.das.ufsc.br/~rabelo/Projects/IFM/Publicacoes/DataMining-SupplyChain.pdf>. Acesso em: 1 nov. 2012.

MACULAN, B. C. M. S. **Manual de normalização**: padronização de documentos acadêmicos do NITEG/UFMG e do PPGCI/UFMG. 2. ed. atual. e rev. Belo Horizonte: UFMG, 2011. Disponível em: <http://www.eci.ufmg.br/normalizacao>. Acesso em: 1 mai. 2015.

MAIMON, Oded; ROKACH, Lior. Introduction to knowledge discovery in databases. In:_____. **The data mining and knwoledge discovery handbook**, [s.l]: Springer, 2005. p. 1-17 . Disponível em: <http://www.ise.bgu.ac.il/faculty/liorr/hbchap1.pdf>. Acesso em: 1 mar. 2014.

MANDA, P.; McCarthy, F.; BRIDGES, S**.** Interestingness measures and strategies for mining multi-ontology multi-level association rules from gene ontology annotations for the discovery of new GO relationships**. J. Biomed. Inform.**, v. 46, n. 5, p. 849-856, oct., 2013. Disponível em: < http://www.ncbi.nlm.nih.gov/pubmed/23850840 > Acesso em: 1 mar.2014.

MARINICA,C.; GUILLET,F.; BRIAND, H. **Post-Processing of Discovered Association Rules Using Ontologies**. [S.l.: s.n.], [ca. 2000] . Disponível em: <http://arxiv.org/ftp/arxiv/papers/0910/0910.0349.pdf>. Acesso em: 1 mar. 2014.

MARINICA, C. **Association Rule Interactive Post-processing using Rule Schemas and Ontologies - ARIPSO**. 2010. Thesis – Department of Computer Science, Ecole poltechnique de l'Universite de Nantes, Nantes, FR. 2010. Disponível em: <http://www.claudiamarinica.com/publications.html>. Acesso em: 1 mar. 2014.

MEALY, G. H. Another look at data. In: AFIPS CONFERENCE, 31th, 1967, Washington. **Proceedings.** Washington: [s.n.], 1967. p. 525-534.

LIMA, J., CARVALHO, C. 2005. **Ontologias**: OWL (Web Ontology Language).Goiás: Universidade Federal de Goiás, 2005. 24 p. Relatório técnico. Disponível em: <http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_004-05.pdf>. Acesso em: 1 mar. 2014.

NIGRO, H. O.; CISARO S. G.; XODO, D. H. **Data mining with ontologies**: implementations, findings and frameworks, information science reference. Hershey: IGI Publishing, 2007.

NOY, N.; MCGUINESS, D. **Ontology Development 101**: a guide to creating your first ontology. Stanford: Stanford University, 2001. Disponível em: <http://www.ksl.stanford.edu/KSL_Abstracts/KSL-01-05.html>. Acesso em: 1 out. 2013.

ONTOLOGY for biomedical investigations. Disponível em: <http://obi-ontology.org/page/Main_Page>. Acesso em: 1 mar. 2014.

ONTOLOGY for general medical science. Disponível em: <http://www.obofoundry.org/cgi-bin/detail.cgi?id=OGMS>. Acesso em: 1 mar. 2014.

OBO Foundry. Disponível em: <http://www.obofoundry.org/>. Acesso em: 1 mar. 2014.

GitHub. Pellet. Disponível em: <https://github.com/complexible/pellet>. Acesso em: 1.fev.2015.

PESQUITA, C. *et al.* **Semantic similarity in biomedical ontologies**, **PLoS Comput. Biol.**, v. 5, n.7, [S.l.: s.n.], 2009. Disponível em: < http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000443> Acesso em: 1 fev.2015.

PROTEGÉ. Disponível em: <http://protegewiki.stanford.edu/wiki/Main_Page>. Acesso em: 1 mar. 2014.

QUONIAM, L. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. Ciência da Informação, Brasília, v.30, n.2, maio/ago. 2001.Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19652001000200004>. Acesso em: 10 out. 2012.

RESNIK, P. Semantic similarity in a taxonomy: an information based measure and its application to problems of ambiguity in natural language. **Journal of Artificial Intelligence Research**, v. 11, p. 95–130, 1999.

RIBEIRO, Lamark dos Santos. Uma abordagem semântica para seleção de atributos no processo de KDD. 2010. 121 f. Dissertação (Mestrado em Informática) - Universidade Federal da Paraíba, João Pessoa, 2010.Disponível em: < http://www.ppgi.di.ufpb.br/?p=1691> Acesso em: 1 out. 2013.

SANTOS, E. C. V.; TSUNODA, D. F.; SILVA, H. F. N. Levantamento bibliográfico referente à mineração de dados como ferramenta de apoio a descoberta ou gestão do conhecimento. In: CONGRESSO BRASILEIRO DE GESTÃO DO CONHECIMENTO. 12.,2012. **Anais**... São Paulo: Universidade Federal da Paraíba, 2012. Disponível em: <http://www.sbgc.org.br/sbgc/kmbrasil-2012/anais/pdf/TC38.pdf>. Acesso em: 1 mar. 2014.

SAÚDE web. [S.l.: s.n.], [ca. 2000]. Disponível em: <www.saudeweb.com.br>. Acesso em: 1 nov. 2012.

SCHEUERMANN, R. H.; WERNER, C.; SMITH, B. Toward an ontological

treatment of disease and diagnosis. 2009. **Translat. Bioinforma**, [S.l], 2009. Disponível em: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3041577/>. Acesso em: 1 mar. 2014.

SCHULZ, S.; JANSEN, L. Formal ontologies in biomedical knowledge representation. **Yearb Med Inform**. v.8, n.1, p.132-146, 2013.

SMITH, B. **Ontology and information systems**. [S.l.: s.n.], 2004. Disponível em: <http://www.ontology.buffalo.edu/ontology>. Acesso em: 10 out. 2013.

SYSTEMATIZED nomenclature of medicine--clinical terms. Disponível em: <http://www.ihtsdo.org/snomed-ct>. Acesso em: 1.mar. 2014.

STERN ,S. D. C; CIFU, A. S.; ALTKORN , D. **Do sintoma ao diagnóstico**: um guia baseado em evidências. Rio de Janeiro: Editora Guanabara Kogan, 2007.

THE infectious disease ontology. Disponível em:<http://infectiousdiseaseontology.org/page/Main_Page>. Acesso em: 1 mai. 2014.

TVERSKY, A. Features of similarity. **Psychological Review**, [S.l], v. 84, n. 4, p. 327–352, 1977.

University of Michigan Medical School. Ontofox. Disponível em:<http://ontofox.hegroup.org/>. Acesso em: 1 mai. 2014.

USCHOLD, M.; GRUNINGER, M. Ontologies: principles, methods and applications. **Knowledge Engineering Review**, v.11, n. 2, p. 93-136, 1996.

VAVPETIC, A., LAVRAC, N. Semantic subgroup discovery systems and workflows in the SDM-Toolkit. **The Computer Journal**, [S.l], 2012. Disponível em: < http://comjnl.oxfordjournals.org/content/early/2012/06/04/comjnl.bxs057>. Acesso em: 1 maio 2014.

VICKERY, B.C. Ontologies. **Journal of Information Science,** v. 23. n. 4, p. 227-86, 1997.

VIDIGAL, F. **Inteligência competitiva**: mapeamento de metodologias de uso estratégico da informação em organizações brasileiras. 2011. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2011.

VIVACQUA, A. S; GARCIA, A. C. B. Mineração de dados baseada em ontologia. In: SEMINÁRIO DE PESQUISA EM ONTOLOGIA NO BRASIL, 2008, Rio de Janeiro. **Anais...** Rio de Janeiro: Universidade Federal Fluminense, 2008, 6 p. Disponível em: <http://www.addlabs.uff.br/Novo_Site_ADDLabs/images/documentos/publicacoes/publicacoes_pdf/trabalhos_anais_congresso/2008/20130809155049_2008%20-%20MINERAO%20DE%20DADOS%20BASEADA%20EM%20ONTOLOGIA.pdf.

> Acesso em: 1 out. 2013.

YOKOME, E.A. Uma ontologia para inserir conhecimento humano em ferramentas de mineração de dados. 2011. 151f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Ciências Exatas e da Natureza, Universidade Metodista de Piracicaba, Piracicaba, 2011. Disponível em: <https://www.unimep.br/phpg/bibdig/aluno/visualiza.php?cod=740> Acesso em: 1 out. 2013.

WAND, Yair; TOREY, Veda C. S; WEBER ,Ron. An ontological analysis of the relationship construct in conceptual modeling. **ACM Transactions on Database Systems**, v. 24, n. 4, p. 494–528, dec. 1999.Disponível em: <http://mba.eci.ufmg.br/downloads/recol/p494-wand.pdf>. Acesso em: 1 out. 2013.

W3C. Web Ontology Language: reference. 2004. Disponível em: <http://www.w3.org/TR/owl-ref//>. Acesso em: 1 out. 2013.

W3C. Simple Knowledge Organization System: reference. 2009.SKOS. Disponível em: <http://www.w3.org/TR/skos-reference/>. Acesso em: 1 out. 2013.

APÊNDICE A – Aquisição de conhecimento

QUESTIONÁRIO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Tema: Modelo para suporte à descoberta de conhecimento em base de dados (KDD): aplicação em estratégias de venda no mercado de medicina diagnóstica
Elaborado por: Lucélia Branquinho
Contato: luceliabranquinho@gmail.com

Por intermédio deste, venho convidá-lo a participar desta pesquisa, de fins acadêmicos em nível de mestrado, da Universidade Federal de Minas Gerais. Esta pesquisa apresenta um estudo sobre a adoção de Sistema de Organização do Conhecimento (SOC), no caso desta pesquisa, o uso de ontologias, para aprimorar a recuperação de informação em base de dados.

Busca resolver o seguinte problema: Obter maior precisão na extração de padrões de regras de associação de exames utilizando conhecimento prévio sobre o assunto. A pesquisa tem como objetivo geral criar uma ferramenta que filtrará os padrões mais relevantes obtidos após mineração de dados por regras de associação utilizando ontologia da área Biomédica. Para elaboração deste trabalho foi definido como amostra os exames relacionados ao diagnóstico de hepatites virais. Portanto, faz necessário validar os conceitos com os especialistas do assunto, ou seja, assessores científicos, especialistas laboratoriais e médicos.

Este questionário tem como objetivo elicitar conhecimento relacionado ao diagnóstico das hepatites virais.

Elicitação de conhecimento pelos especialistas

Para entendimento acerca do roteiro de entrevista de elicitação de conhecimento, cabe esclarecer alguns conceitos utilizados como referência. Os termos descritos, compõem o "quadro terminológico", que abrange as doenças, manifestações, testes laboratórios entre outras entidades relacionadas ao modo como as doenças são reconhecidas e interpretadas pela medicina (Scheuermann et al, 2009).

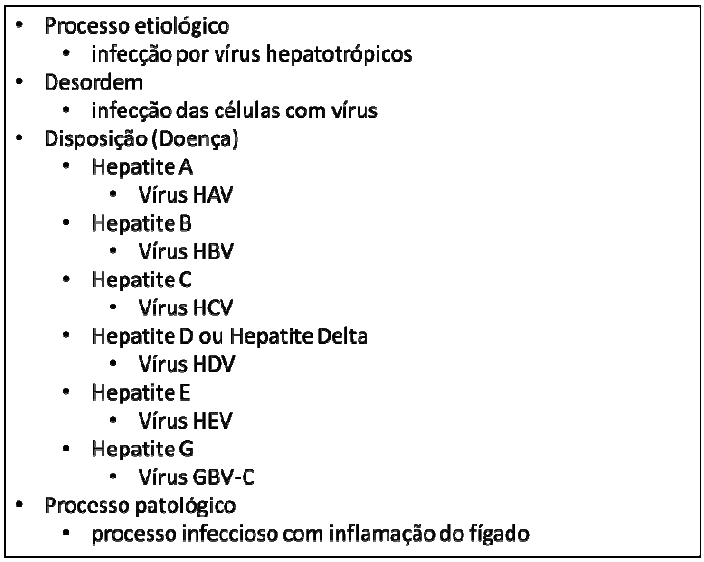
Na fase chamada processo etiológico, considera-se que há um corpo humano saudável, com características normais de acordo com parâmetros médicos. Como resultado desse processo, ocorre uma mudança física no indivíduo, dando origem à doença. Na manifestação pré-clínica da doença, o corpo desenvolve desordens, que são portadoras de disposições. A fase de curso da doença começa com a manifestação da doença. Nesse momento, a desordem se manifesta por meio de sintomas e sinais (manifestações clinicas), sendo o primeiro aqueles

que os pacientes conseguem sentir ou identificar e o segundo, que é determinado pelo médico através de exames físicos e testes laboratoriais.

Na fase resposta terapêutica, uma amostra do paciente é obtida de alguma parte do corpo, a fim de submetê-la a testes laboratoriais. Esses resultados são registrados no prontuário médico como um quadro clínico. O quadro clínico é interpretado pelo médico na busca pelo diagnóstico. Nesse ponto, é possível que se estabeleça um plano de tratamento, baseado neste diagnóstico, de modo que o corpo possa retornar à normalidade. O plano é o resultado do diagnóstico fundamentado no processo de interpretação do quadro clínico. O quadro clínico é composto por registros da representação de sintomas e sinais, bem como os resultados do exame físico e laboratorial (Scheuermann *et al.,* 2009).

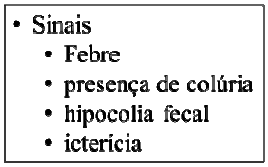
Com base neste quadro terminológico, propõe-se a elicitação do conhecimento sobre hepatites virais através das questões relacionadas abaixo:

1.



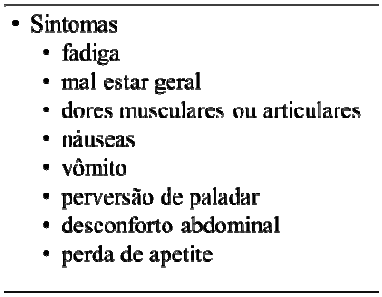
A organização está correta? Esses conceitos se aplicam à todas as hepatites virais? Caso negativo, descreva suas considerações.

2.



São esses os possíveis SINAIS das hepatites virais ? Existem outros?

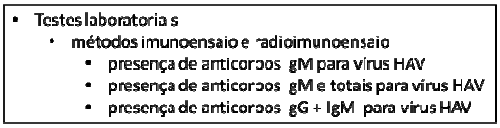
3.



São esses os possíveis SINTOMAS das hepatites virais? Existem outros?

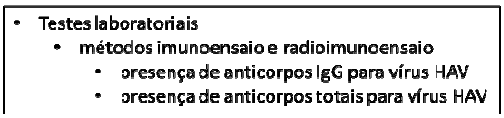
As questões de 4 até 21 relacionam os testes laboratoriais considerando as especificidades de cada tipo de hepatite viral.

4.



São esses os possíveis exames laboratoriais com marcadores sorológicos solicitados para identificação da infecção pelo vírus HAV na fase aguda ou crônica? Existem outros?

5.

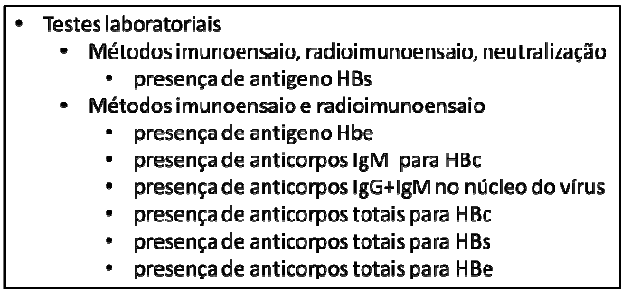


São esses os exames laboratoriais sorológicos solicitados para acompanhamento da evolução clínica ou comprovação de imunização ao vírus HAV? Existem outros?

6.

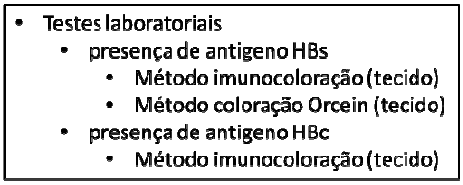
São esses os exames laboratoriais moleculares solicitados para identificação do vírus HAV? Existem outros?

7.



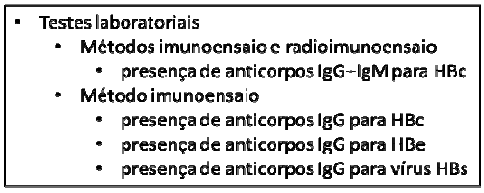
São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para identificação da infecção pelo vírus HBV na fase aguda ou crônica? Existem outros?

8.



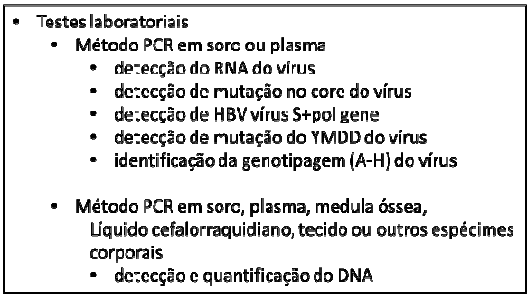
São esses os possíveis exames laboratoriais não sorológicos solicitados para identificação da infecção pelo vírus HBV na fase aguda ou crônica? Existem outros?

9.



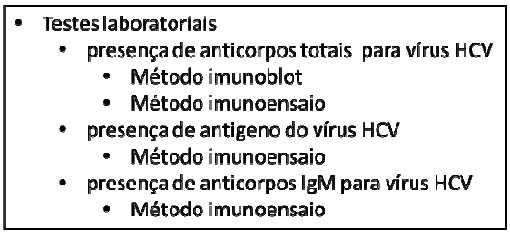
São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para acompanhamento da evolução clínica da infecção pelo vírus HBV ou comprovação de imunização ao vírus? Existem outros?

10.



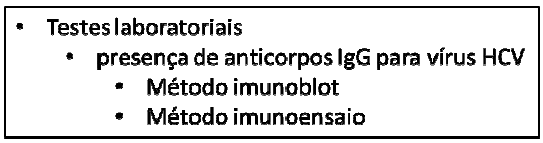
São esses os exames laboratoriais moleculares solicitados para identificação do vírus HBV? Existem outros?

11.



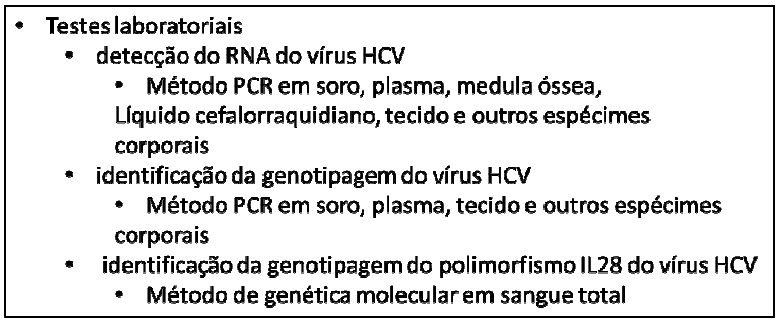
São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para identificação da infecção pelo vírus HCV na fase aguda ou crônica? Existem outros?

12.



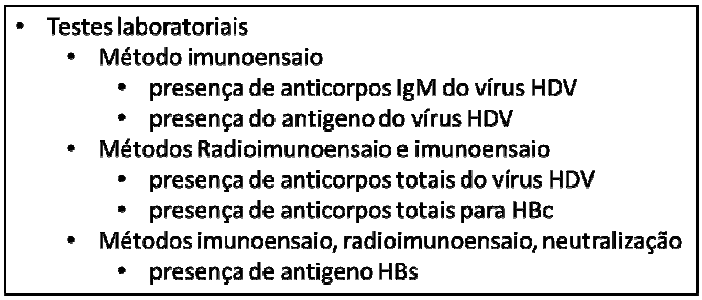
São esses os exames laboratoriais de marcadores sorológicos solicitados para comprovação de imunidade ao vírus HCV e acompanhamento da evolução clínica ? Existem outros?

13.



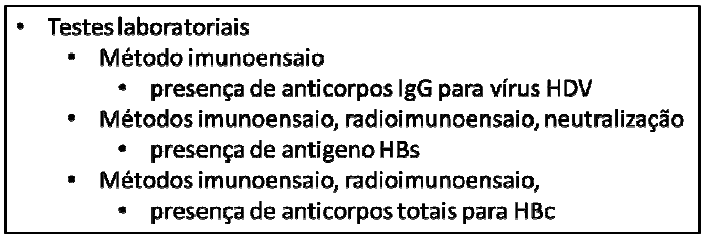
São esses os exames laboratoriais moleculares solicitados para identificação do vírus HCV? Existem outros?

14.



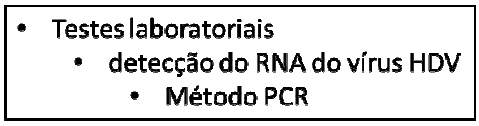
São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para identificação da infecção pelo vírus HDV na fase aguda ou crônica? Existem outros?

15.



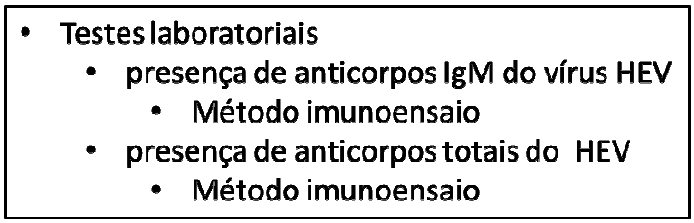
São esses os exames laboratoriais de marcadores sorológicos solicitados para comprovação de imunidade ao vírus HDV ou acompanhamento da evolução clínica? Existem outros?

16.



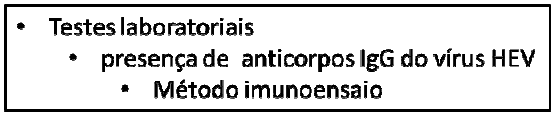
São esses os exames laboratoriais moleculares solicitados para identificação do vírus HDV? Existem outros?

17.



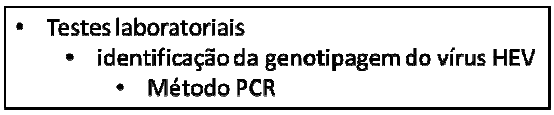
São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para identificação da infecção pelo vírus HEV na fase aguda ou crônica? Existem outros?

18.



São esses os exames laboratoriais sorológicos solicitados para comprovação de imunidade ao vírus HEV ou acompanhamento da evolução clínica? Existem outros?

19.

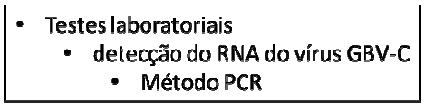


São esses os exames laboratoriais moleculares solicitados para identificação do vírus HEV? Existem outros?

20.

São esses os possíveis exames laboratoriais de marcadores sorológicos solicitados para identificação da infecção pelo vírus GBV-C na fase aguda ou crônica? Existem outros? Quais os métodos utilizados para realização do exame?

21.



São esses os exames laboratoriais moleculares solicitados para identificação do vírus GBV-C? Existem outros?

APÊNDICE B – Entrevista validação

MESTRADO EM CIÊNCIA DA INFORMAÇÃO
UNIVERSIDADE FEDERAL DE MINAS GERAIS

Hepatites virais (A,B,C,D,E,G)

Regras de associação
Validação com os especialistas

Tema: Modelo para suporte à descoberta de conhecimento em base de dados (KDD): aplicação em estratégias de venda no mercado de medicina diagnóstica

1

Regras de associação

A técnica de **mineração** por **regras de associação** busca correlação entre conjuntos de **itens frequentes** em uma série de dados ou transações. A partir de conjuntos de elementos que aparecem juntos com pelo menos alguma frequência (suporte), as regras de associação representam a condição "SE antecedente ENTÃO consequente" com garantia probabilística (confiança), que sempre que o antecedente ocorrer o consequente também estará presente. (FERRAZ, 2008).

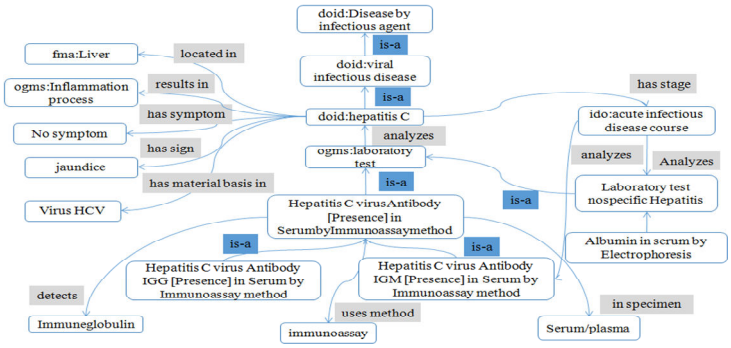
O suporte de uma regra P → Q é definido como o percentual de transações da base de dados em que o antecedente e consequente da regra aparecem na mesma transação

Para mensurar a **dependência entre os itens** é utilizada a **métrica Lift** (Gonçalves, 2011). É uma métrica utilizada para avaliar dependências entre antecedente e consequente, quanto maior o valor do lift, mais interessante a regra, pois A aumentou (“lifted”) B numa maior taxa.

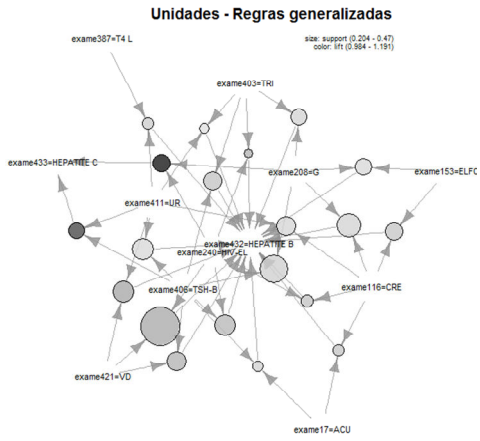
2

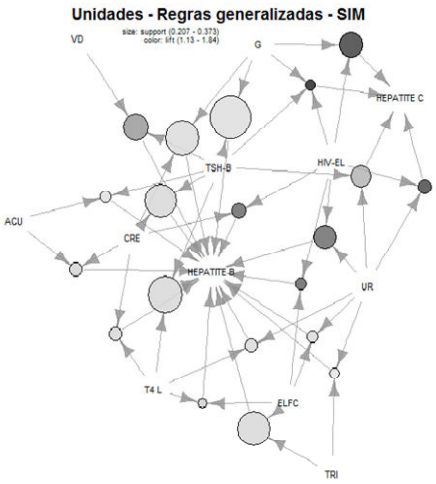
Fonte: Scheuermann et al, 2009

Modelagem do conhecimento

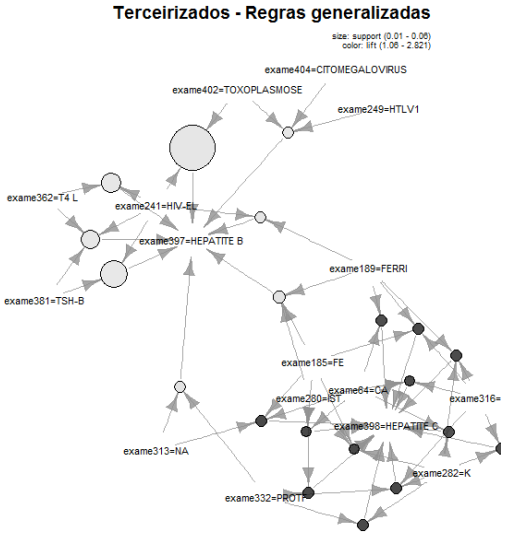


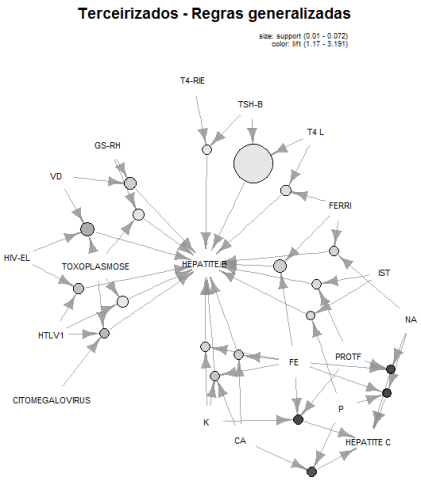
3



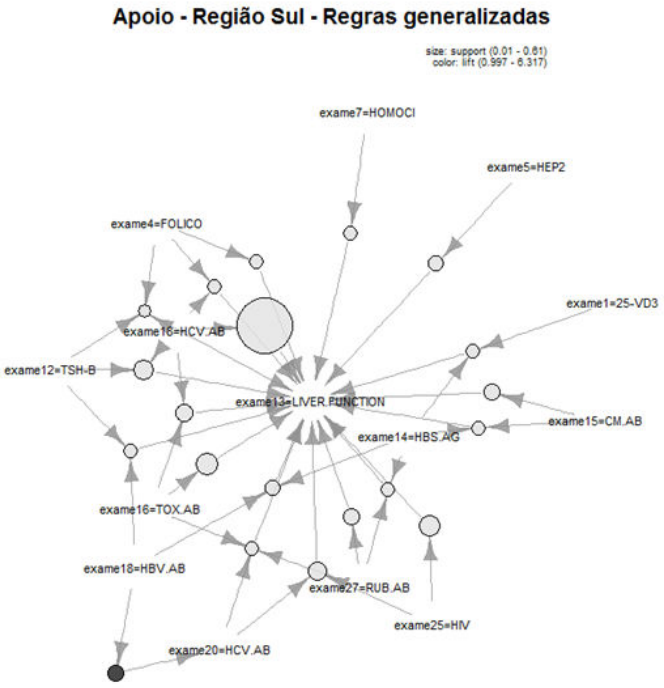


Considerando que o suporte é representando pelo tamanho da bola e a dependência entre os itens (lift) pela cor. Podemos afirmar que as regras mais relevantes são as apresentadas no gráfico a esquerda com a cor mais escura?





Considerando que o suporte é representando pelo tamanho da bola e a dependência entre os itens (lift) pela cor. Podemos afirmar que as regras mais relevantes são as apresentadas no gráfico a esquerda com a cor mais escura?

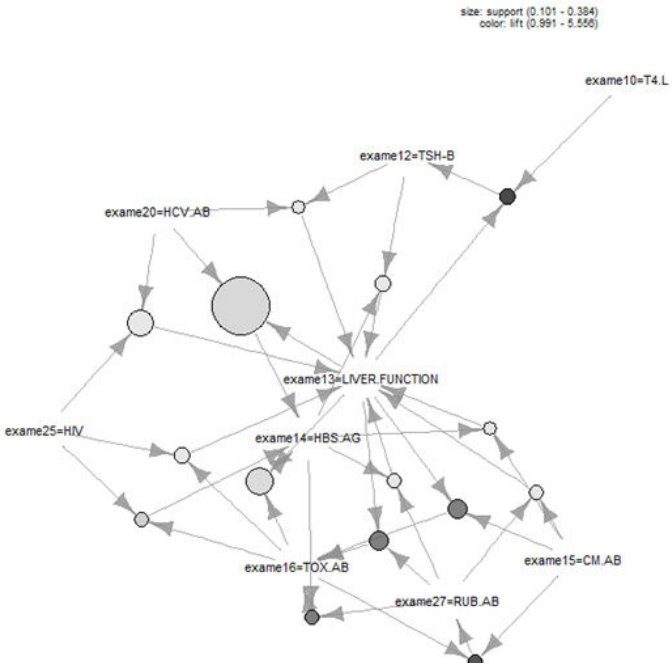


Regras classificadas

Antecedente		Consequente
HCV.AB	HIV	HBS.AG
LIVER.FUNCTION	HBV.AB	HCV.AB
HBS.AG	HCV.AB	LIVER.FUNCTION
HAV.AB	HCV.AB	LIVER.FUNCTION
LIVER.FUNCTION	HIV	HBS.AG
HCV.AB	HIV	LIVER.FUNCTION
TOX.AB	HCV.AB	LIVER.FUNCTION
HBS.AG	TOX.AB	LIVER.FUNCTION
TSH-B	HCV.AB	LIVER.FUNCTION
T4.L	LIVER.FUNCTION	TSH-B

Podemos afirmar que a melhor classificação das regras considerando a semelhança entre as doenças, sintomas e local de ocorrência são as apresentadas na tabela?

Apoio - Região Nordeste - Regras generalizadas



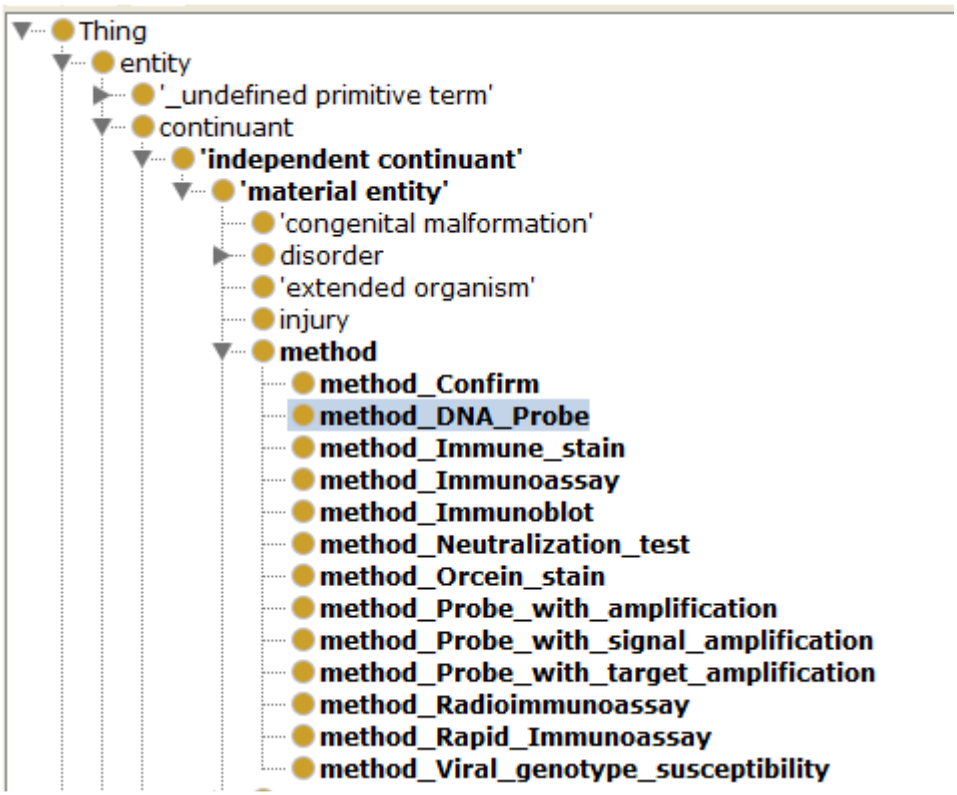
Regras classificadas

Antecedente			Consequente
HCV.AB	HIV		HBS.AG
LIVER.FUNCTION	HCV.AB		HBS.AG
LIVER.FUNCTION	HIV		HBS.AG
HCV.AB	HIV		LIVER.FUNCTION
LIVER.FUNCTION	TOX.AB	HIV	HBS.AG
HBS.AG	CM.AB		LIVER.FUNCTION
HBS.AG	RUB.AB		LIVER.FUNCTION
HBS.AG	RUB.AB		TOX.AB
CM.AB	RUB.AB		LIVER.FUNCTION
TOX.AB	HIV		HBS.AG
LIVER.FUNCTION	TOX.AB		HBS.AG
LIVER.FUNCTION	RUB.AB		TOX.AB
LIVER.FUNCTION	CM.AB		TOX.AB
TOX.AB	HIV		LIVER.FUNCTION
CM.AB	TOX.AB		RUB.AB
TSH-B	HCV.AB		LIVER.FUNCTION
TSH-B	HBS.AG		LIVER.FUNCTION
T4.L	LIVER.FUNCTION		TSH-B

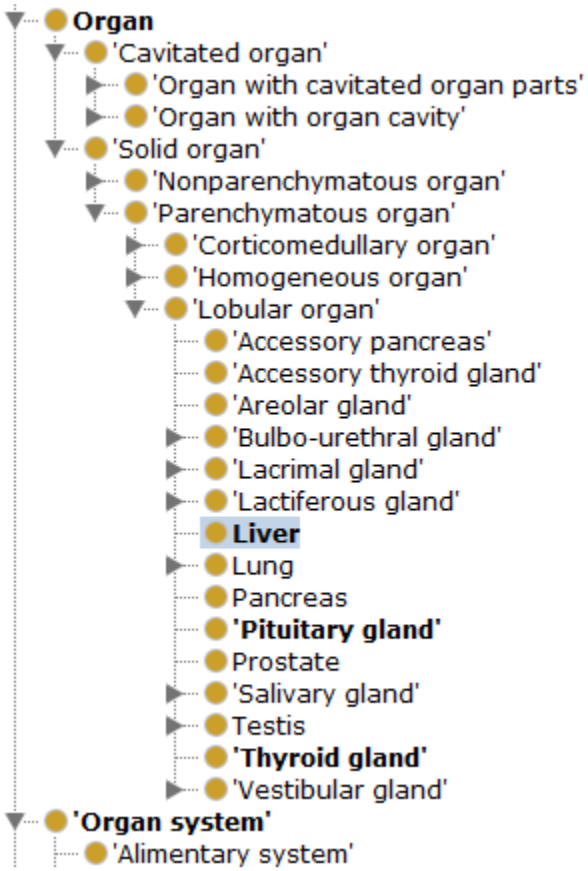
Podemos afirmar que a melhor classificação das regras considerando a semelhança entre as doenças, sintomas e local de ocorrência são as apresentadas na tabela?

APÊNDICE C – HVO

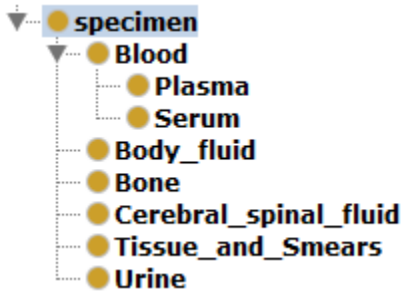
1) Extrato da ontologia criada a partir dos conceitos avaliados do LOINC para representar o procedimento utilizado para fazer a medição ou observação do teste complementar



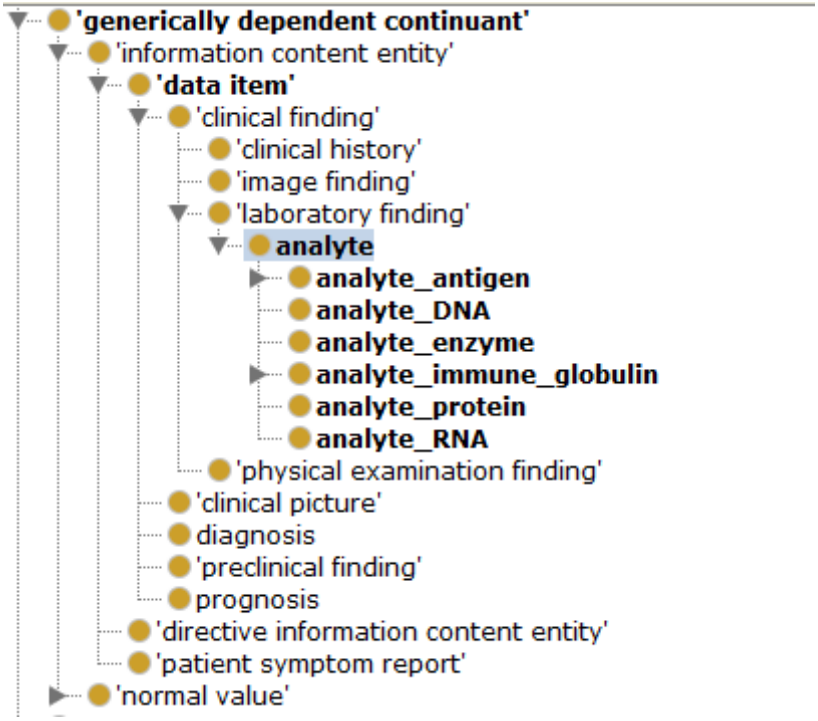
2) Extrato da ontologia FMA para representar os órgãos e sistemas do corpo humano



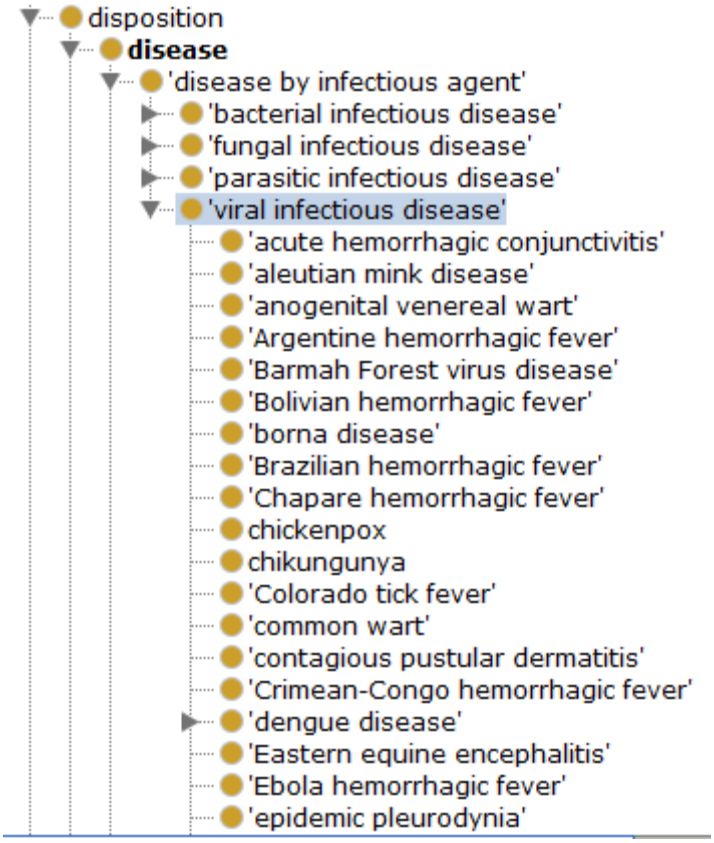
3) Extrato da ontologia criada a partir dos conceitos avaliados do LOINC para representar o contexto ou espécime tipo no qual a observação foi feita (exemplo: sangue, urina, ...)

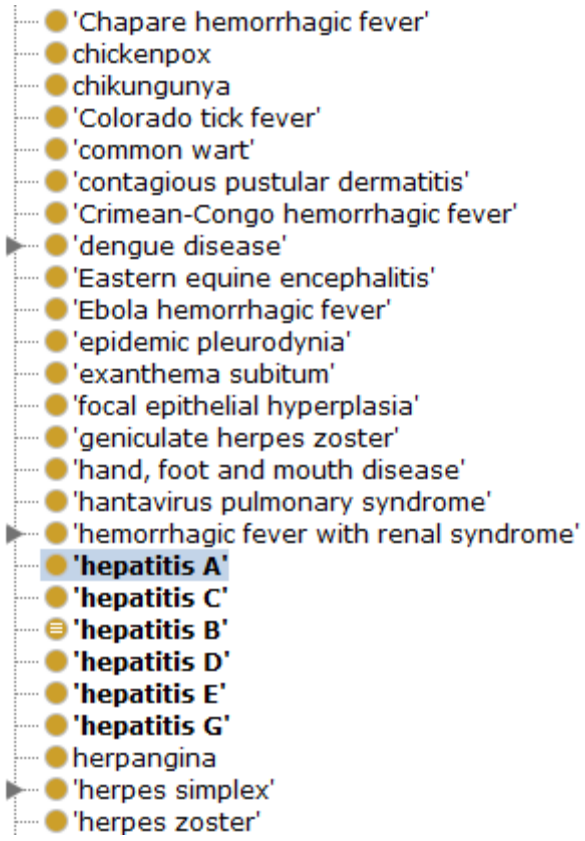


4) Extrato da ontologia criada a partir dos conceitos avaliados do LOINC para representar a substância química presente numa amostra e cuja concentração se pretende determinar



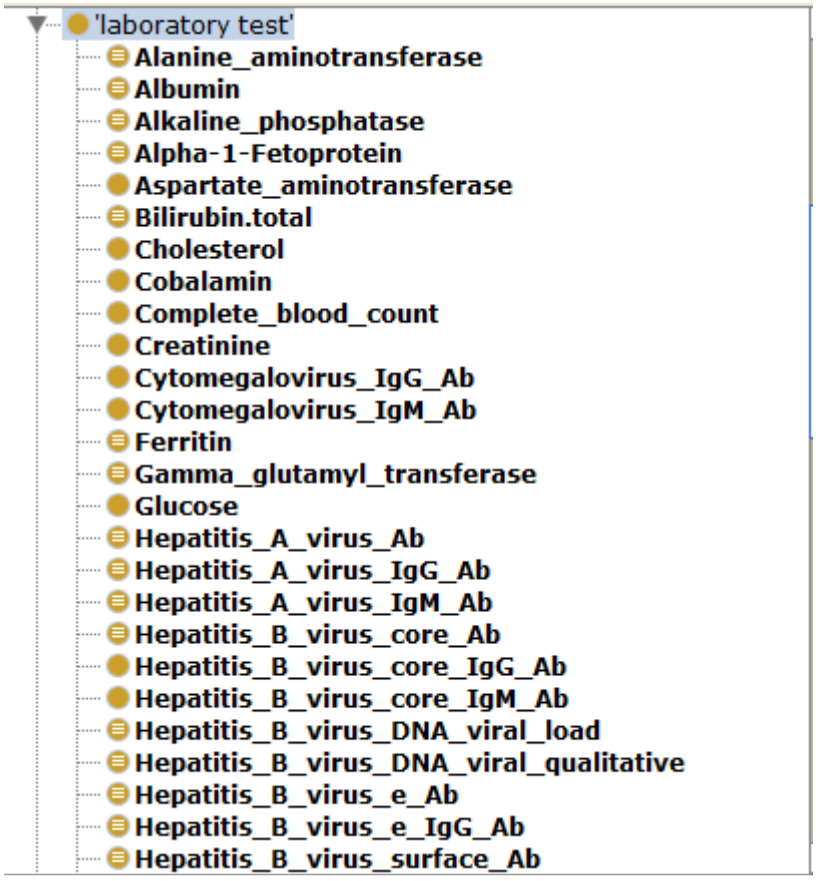
4) Extrato da ontologia baseada na representação das doenças na ontologia DOID



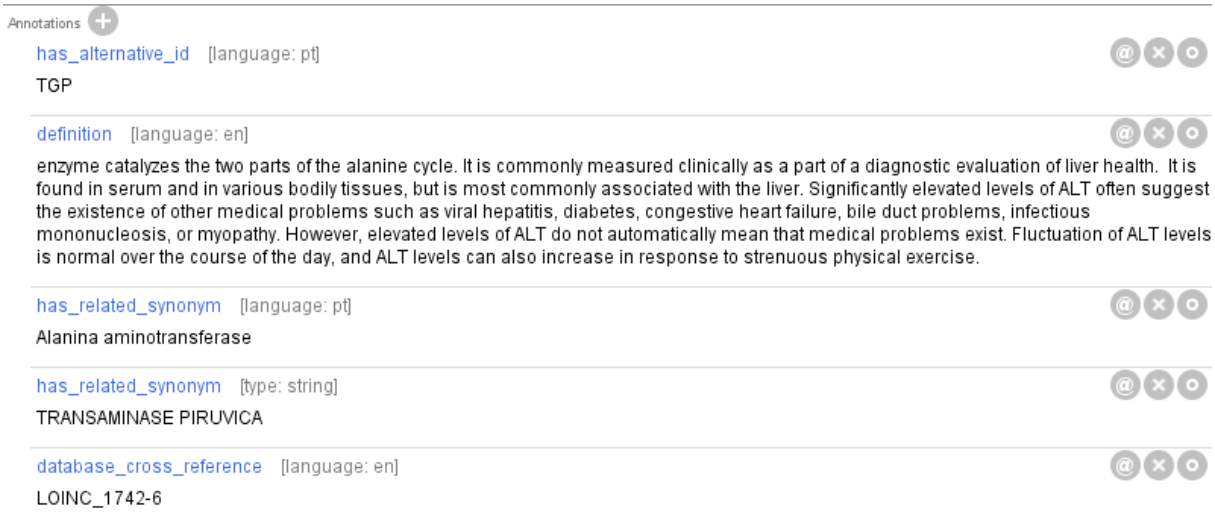


5) Extrato da ontologia baseada nas anotações sobre as doenças descrita na ontologia DOID

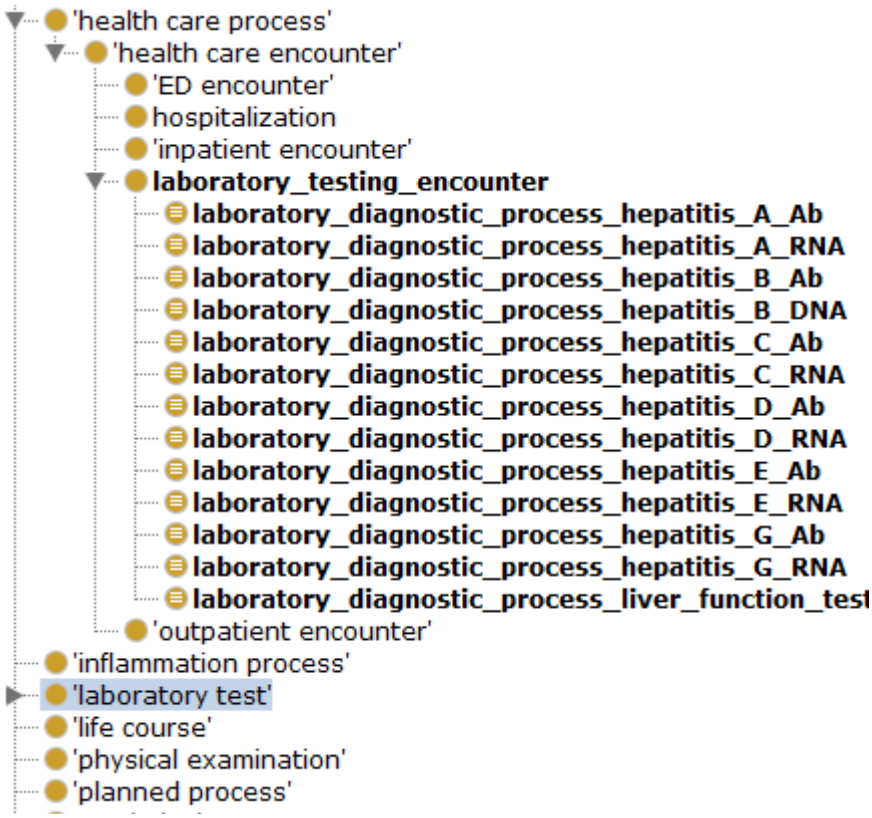
6) Extrato da ontologia criada a partir dos conceitos avaliados do LOINC para os testes complementares



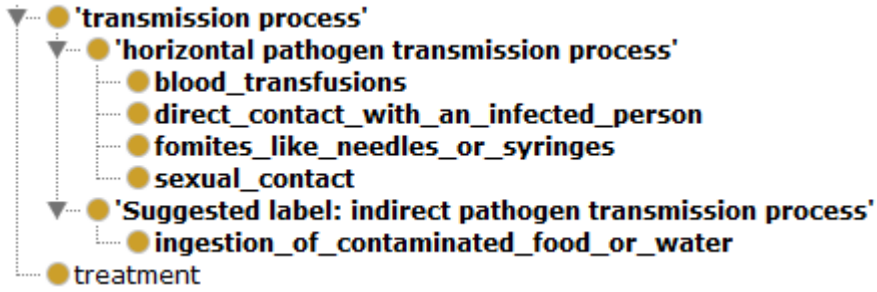
7) Extrato da ontologia baseada nas anotações do LOINC sobre os testes complementares



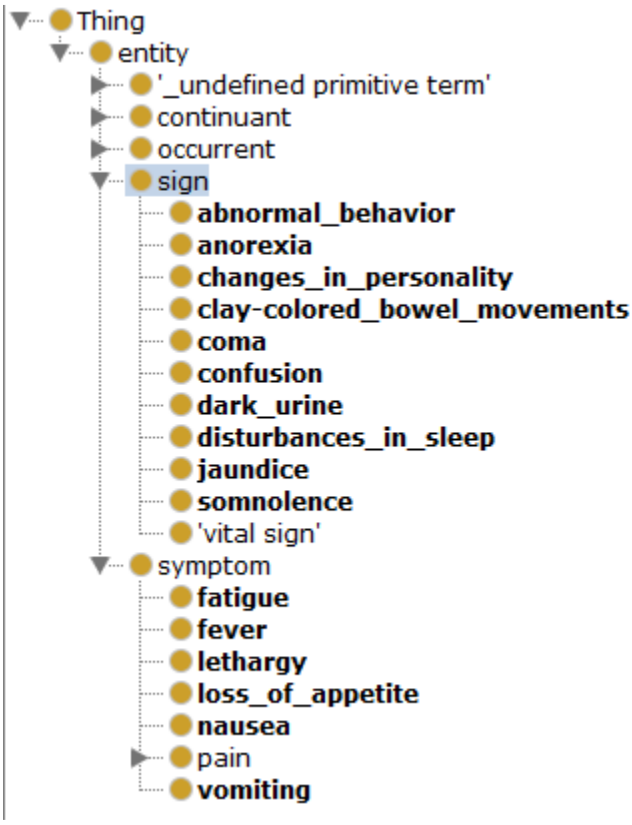
8) Extrato da ontologia HVO para representar a prescrição médica dos testes complementares para diagnóstico de hepatites virais



9) Extrato da ontologia IDO para representar o processo de transmissão da doença



10) Extrato da ontologia para representar os sintomas e sinais relacionados as doenças


Thing
entity
'_undefined primitive term'
continuant
occurrent
sign
abnormal_behavior
anorexia
changes_in_personality
clay-colored_bowel_movements
coma
confusion
dark_urine
disturbances_in_sleep
jaundice
somnolence
'vital sign'
symptom
fatigue
fever
lethargy
loss_of_appetite
nausea
pain
vomiting

APÊNDICE D – Axiomas de equivalência

1) Axioma de equivalência Hepatite A Anticorpos

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

and (diagnoses only 'hepatitis A')))

and (is_composed_of min 0 ('laboratory test'

and (diagnostic_evaluation some Liver)))

2) Axioma de equivalência Hepatite B

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

and (diagnoses only 'hepatitis B')))

and (is_composed_of min 0 ('laboratory test'

and (diagnostic_evaluation some Liver)))

3) Axioma de equivalência Hepatite C

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

and (diagnoses only 'hepatitis B')))

and (is_composed_of min 0 ('laboratory test'

and (diagnostic_evaluation some Liver)))

4) Axioma de equivalência Hepatite D

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

and (diagnoses only 'hepatitis D')))

and (is_composed_of min 0 ('laboratory test'

and (diagnostic_evaluation some Liver)))

5) Axioma de equivalência Hepatite E

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

and (diagnoses only 'hepatitis E')))

and (is_composed_of min 0 ('laboratory test'

and (diagnostic_evaluation some Liver)))

6) Axioma de equivalência Hepatite G

laboratory_testing_encounter

and (is_composed_of some

('laboratory test'

```
    and (diagnoses only 'hepatitis G')))
and (is_composed_of min 0 ('laboratory test'
and (diagnostic_evaluation some Liver)))
```