UNIVERSIDADE FEDERAL DE SANTA CATARINA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA E GESTÃO DO CONHECIMENTO

Rafael Andrade

Um modelo para recuperação e comunicação do conhecimento em documentos médicos

Tese de doutorado submetida à Universidade Federal de Santa Catarina como parte dos requisitos para a obtenção do grau de doutor em Engenharia e Gestão do Conhecimento
Orientador: Prof. Dr. rer. nat. Aldo von Wangenheim
Coorientador: Prof. Dr. Mario Dantas

Florianópolis

2011

Catalogação na fonte pela Biblioteca Universitária
da
Universidade Federal de Santa Catarina

A553m Andrade, Rafael
Um modelo para recuperação e comunicação do conhecimento em documentos médicos [tese] / Rafael Andrade ; orientador, Aldo von Wangenheim. - Florianópolis, SC, 2011.
180 p.: il., grafs., tabs.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Tecnológico. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Inclui referências

1. Engenharia e gestão do conhecimento. 2. Ontologia. 3. Web semântica. 4. Sistemas de recuperação da informação - Modelos. I. Wangenheim, Aldo von. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. III. Título.

CDU 659.2

Rafael Andrade

UM MODELO PARA RECUPERAÇÃO E COMUNICAÇÃO DO CONHECIMENTO EM DOCUMENTOS MÉDICOS

Esta Tese foi julgada adequada para a obtenção do título de Doutor em Engenharia e Gestão do Conhecimento Área de Concentração Engenharia do Conhecimento e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Florianópolis, 03 de março de 2011.

_____

Prof. Paulo Maurício Selig, Dr.,
Coordenador do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento.

Banca Examinadora

_____

Prof. Aldo von Wangenheim, Dr. rer. nat., UFSC – Presidente - Orientador

_____

Prof. Mario Dantas, Dr., UFSC - Coorientador

_____

Prof. Renato Fileto, Dr., UFSC – Moderador

_____

Prof. Luiz Felipe Nobre, Dr., UFSC - Membro

_____

Prof. Eros Comunello, Dr., UNIVALI - Membro

_____

Prof[a]. Francis Solange Vieira Tourinho, Dr.[a], UFRN – Membro

AGRADECIMENTOS

A Deus, por iluminar meu caminho e me dar forças para seguir sempre em frente.

Ao professor Aldo von Wangenheim, pela amizade, pela sua competência em exercer seu papel como orientador e pela grande ajuda em proporcionar meu afastamento para o doutorado sanduíche na Alemanha.

Ao Professor Mário Dantas, coorientador, pelos conselhos, discussões e contribuições sobre o tema.

Ao professor Luiz Felipe Nobre, que contribuiu com seu conhecimento de especialista médico na área de telemedicina.

A *Ruprecht-Karls-Universität Heidelberg* na Alemanha, em especial ao meu professor alemão, Dr. Hans-Peter Meinzer por proporcionar um excelente ambiente de trabalho, companheirismo, compreensão sobre meu pouco conhecimento na língua alemã, pela ajuda e incentivo no aprendizado da língua e, por proporcionar uma integração com os colegas de trabalho de modo que eu me sentisse o mais próximo de casa possível.

A todos os colegas de trabalho do *German Cancer Research Center* na Alemanha, pela compreensão, companheirismo e amizade, durante minha estada na Alemanha.

Ao Rodrigo B. Cabral, ao Cloves L. Barcellos Jr. e ao Fernando Bertoldi (o Pavezinho), que muito contribuíram para a implementação do protótipo.

Ao grande amigo Douglas Macedo, pelo incentivo e por não deixar que eu desistisse no meio do caminho.

Aos meus familiares, que mesmo acompanhando minha luta de longe, por sempre apoiarem meus estudos, em especial à minha mãe, Tânia e ao meu pai Hilário, por sempre acreditarem em mim.

Em especial, à minha esposa Sabrina, por apoiar cada passo, sempre ao meu lado com carinho, companheirismo, otimismo e incentivo.

E por fim, a todos os colegas de trabalho do Grupo Cyclops que de uma forma ou de outra contribuíram para o desenvolvimento desse trabalho.

RESUMO

O grande número de informações disponíveis, que estão em diferentes fontes de dados, exige cada vez mais processamento dos motores de busca. Recuperar informações que estão nessas bases de dados com a melhor precisão possível é um dos desafios a serem alcançados dentro do contexto desta tese. Os registros clínicos médicos contêm uma imensa gama de informações, normalmente escritas em forma de texto livre e sem um padrão linguístico. Os médicos não escrevem os diagnósticos e os laudos do paciente com o uso de elementos de estilo, o que dificulta o processamento e a recuperação da informação por parte dos sistemas computacionais. Consequentemente, obter o conhecimento a partir desses dados não é uma tarefa fácil para um motor de busca. Este trabalho apresenta o desenvolvimento de um modelo, que permite recuperar o conhecimento de informações textuais em documentos médicos. Técnicas de expansão de pesquisas, que utilizam detecção de ativos de conhecimento da ontologia DeCS e de dicionários linguísticos, são utilizadas. O objetivo é ampliar o universo de pesquisa do usuário e criar uma base de conhecimento para permitir o seu reúso. A proposta de tese aqui apresentada difere dos anteriores porque a intenção é retornar às pesquisas dos usuários uma série de documentos médicos muito mais eficazes do que nas tradicionais ferramentas de busca. Com o intuito de melhorar os resultados de uma pesquisa, anotações semânticas e detecção de expressões negativas serão utilizadas para processar os textos médicos. O estudo de caso apresentado no final mostra que, dos dez primeiros resultados do modelo ora proposto, alcançou-se uma média de 90% de precisão, enquanto que o modelo booleano limitou-se a 60%, e com o diferencial de que no modelo tradicional, o usuário teve que refazer suas consultas várias vezes até chegar a um resultado satisfatório, ao passo que no modelo semântico obteve êxito já na primeira consulta. Justamente porque o usuário não encontrou uma resposta nas primeiras pesquisas no modelo booleano, os tempos de resposta médios foram de 49 minutos, contra 0,6 segundos do novo modelo. Conclui-se, dessa forma, que o usuário não precisará despender muito tempo para encontrar a informação ou não precisará procurar em diferentes bases de dados a fim de encontrar a informação necessária

ABSTRACT

The large amount of information available in different data sources has been demanding more power processing from search engines. Retrieving information from these databases with the best possible precision is the great challenge to be achieved. Clinical medical records contain a vast range of information, usually written in free text form and without a standard language. Doctors do not write the diagnosis and patient reports with style elements, thus making it difficult for computer systems to process and retrieve information. Consequently, obtaining knowledge from this data is not an easy task for a search engine. This work presents the development of a model that allows recovering knowledge from textual information in medical documents. Query expansion techniques, which apply knowledge detection assets from the DeCS ontology and language dictionaries, will be used. The goal is to expand the user search universe and to create a knowledge base to allow its reuse. This thesis proposal differs from previous ones because the intention is to respond to the user's queries with a series of medical documents much more effectively than traditional search engines. In order to improve the search results, semantic annotations and negation detection will be used to process medical texts. The case study presented at the end shows that the proposed model was able to achieve a mean accuracy of 90% in its first ten results, while the Boolean model was limited to only 60%, with the difference that in the latter model the user had to restate their queries several times until they could get a satisfactory outcome, while in the former results were successful in the first query. The conclusion is that the user will not need to spend much time to find information or will not need to search in different databases to find the necessary information.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANVISA	Agência Nacional de Vigilância Sanitária
CIT/SC	Centro de Informações Toxicológicas de Santa Catarina
CLEF	Cross Language Evaluation Forum
CID	Classificação Internacional de Doenças
CPGs	Clinical Practice Guidelines
DeCS	Descritores em Ciências da Saúde
ECG	Eletrocardiografia
GC	Gestão do Conhecimento
GO	Genome Annotations
HSDB	Hazardous Substances Data Bank
HU/UFSC	Hospital Universitário da Universidade Federal de Santa Catarina
IHTSDO	Organização Internacional do Desenvolvimento dos Padrões da Terminologia da Saúde
LOINC	Logical Observation Identifiers Names and Codes
MeSH	Medical Subject Headings
NLP	Natural Language Processing
PLN	Processamento de Linguagem Natural
RCTM	Rede Catarinense de Telemedicina
SBC	Sistemas Baseados em Conhecimento
SE	Sistemas Especialistas
SES/SC	Secretaria de Estado da Saúde de Santa Catarina
SNOMED	Systematized Nomenclature of Medicine
SNOP	Systematized Nomenclature for Pathology
STT	Sistema Catarinense de Telemedicna e Telessaúde
SUS	Sistema Único de Saúde Pública
RI	Recuperação da Informação
RM	Ressonância Magnética
TF-IDF	Term Frequency – Inverse Document Frequency
TI	Tecnologia da Informação
TC	Tomografia Computadorizada
UMLS	Unified Medical Language System
US	Ultrassonografia

SUMÁRIO

1 INTERODUÇÃO

1.1 CONTEXTUALIZAÇÃO

Telemedicina é baseada no uso de informações eletrônicas e tecnologias de telecomunicações para oferecer suporte a distância para pacientes, profissionais de saúde, clínicas e hospitais (Mcneill *et al.*, 1998). E é essa disponibilidade de dados a distância que tem despertado grande interesse na telemedicina no mundo, tanto por parte de instituições de saúde, quanto por parte de governos. Esse também é o motivo que levou à disseminação dos sistemas de arquivamento e comunicação de imagens (do inglês, *Picture Archiving and Communication System* - PACS).

O PACS possui como um de seus componentes, a distribuição de imagens e foi originalmente desenvolvido para serviços de radiologia com a finalidade de capturar imagens médicas eletronicamente, ao invés de utilizar filme (Huang, 2003; Clunie, 2008). Atualmente, o PACS não se limita aos serviços de radiologia e vem sendo estendido para outros serviços clínicos de imagens como cardiologia, patologia, dentre outros (Tie, 2009). Essa extensão permitiu que fossem disponibilizados diversos tipos de dados médicos, e com a utilização da telemedicina foi possível também a manipulação e a avaliação remota dessas informações.

O conhecimento adquirido em sistemas de telemedicina permitiu que em 2005 a Universidade Federal de Santa Catarina, juntamente com a Secretaria de Estado da Saúde de Santa Catarina – SES/SC desenvolvesse um sistema de telemedicina para auxílio à saúde do paciente. Nesse projeto estava prevista a ampliação do parque de equipamentos de média e alta complexidade conectados em rede, como por exemplo, eletrocardiogramas, equipamentos de ultra-sonografia (US), de tomografia computadorizada (TC) e de ressonância magnética (RM). Esses equipamentos foram distribuídos pelo Estado de Santa Catarina e o projeto foi chamado de Rede Catarinense de Telemedicina (RCTM).

Estava previsto também o desenvolvimento de um sistema para armazenamento centralizado e distribuição de informações de pacientes, imagens e laudos de imagens diretamente na web (Wallauer *et al.*, 2008). Na segunda etapa desse projeto, a rede foi expandida para também atender aos serviços de administração laboratorial on-line

integrado ao Portal de Telemedicina. O sistema desenvolvido tornou mais ágil o processo de administração, manipulação e visualização dos exames laboratoriais, minimizando a duplicação de tarefas e automatizando o processo de divulgação e registro dos exames laboratoriais realizados pelos laboratórios de análises clínicas do Estado de Santa Cataria.

No ano de 2007, o Ministério da Saúde propôs a criação do Programa de Telessaúde Brasil (http://www.telessaude.org.br). Por ter experiência em Telemedicina, Santa Catarina foi um dos Estados a fazer parte desse programa. O Projeto Nacional de Telessaúde em Apoio à Atenção Primária foi realizado simultaneamente em nove Estados brasileiros, contendo um Núcleo de Telessaúde em cada Estado escolhido. Cada núcleo deveria vincular pelo menos 100 pontos de Telessaúde, preferencialmente em municípios diferentes, totalizando, 900 Pontos de Telessaúde instalados e distribuídos pelo Brasil. Cada um desses pontos de Telessaúde está funcionando dentro das Unidades Básicas de Saúde dos municípios selecionados e buscam atender pelo menos 2.700 Equipes de Saúde da Família em todo o país.

O projeto Telessaúde objetiva contribuir para a qualificação profissional e auxiliar os procedimentos assistenciais da rede de Atenção Primária. As Equipes de Saúde da Família dos municípios credenciados recebem apoio remoto por meio de serviços de segunda opinião formativa (tele-consultorias e suporte a dúvidas clínicas) e ensino a distância (capacitações e disponibilização de material de aprendizagem multimídia diretamente em um sistema web).

Em 2010, esses dois projetos foram unificados e formaram o Sistema Catarinense de Telemedicina e Telessaúde (STT). A partir dessa integração, um único sistema oferece, além do envio de exames e laudos à distância de diversas modalidades, acesso por parte dos pacientes aos seus exames, palestras virtuais aos profissionais de saúde, segunda opinião formativa e também capacitação continuada aos profissionais da Atenção Básica.

A Figura 1 apresenta a página principal de entrada do STT. Podem acessar o sistema, profissionais de enfermagem, médicos que requisitaram os exames, médicos especialistas e o próprio paciente. Maiores detalhes sobre permissões de acesso podem ser vistas no apêndice B.

Figura 1: Sistema Catarinense de Telemedicina e Telessaúde: tela inicial.

O principal objetivo da RCTM foi descentralizar o processo e execução de exames de pacientes atendidos pelo Sistema Único de Saúde Pública (SUS), de forma a diminuir a superlotação de hospitais dos grandes centros, direcionando o paciente ao hospital mais próximo de sua região. Essa filosofia foi implantada com base em uma tecnologia de software de baixo custo operacional e alta eficiência em termos de aproveitamento da mão de obra do médico especialista (Maia *et al.*, 2006).

Todos os exames de média e alta complexidade são enviados ao servidor central. Mesmo quando há um médico especialista em um hospital local para emitir os laudos, os exames necessitam ser enviados

ao servidor central para que se tenha o registro do exame e permitir o acesso online. Dessa forma, o médico especialista pode emitir o laudo do exame do próprio local onde as imagens foram geradas e o paciente pode acessar o resultado de seu exame diretamente na *web*.

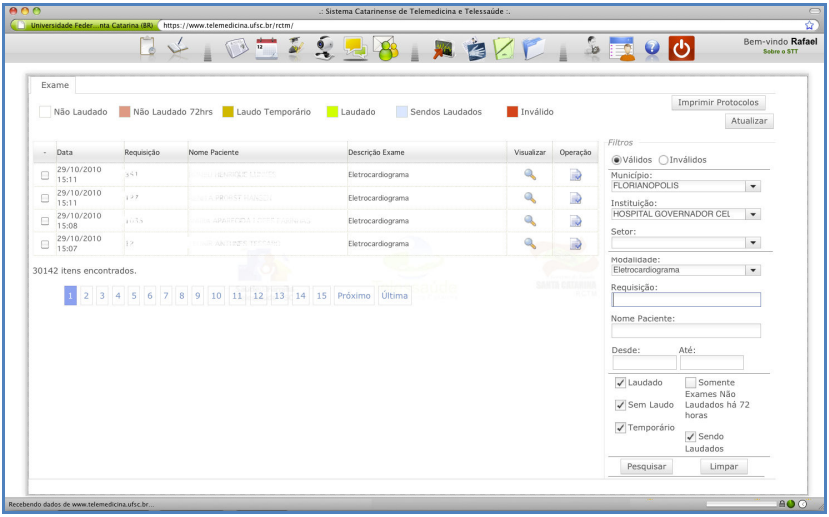


Figura 2: Tela principal do Sistema Catarinense de Telemedicina.

Atualmente esse serviço está presente em 287 dos 293 municípios de Santa Catarina, sendo que 193 já possuem meios para envio de exames a distância. Desde o inicio do funcionamento em 2005, até março de 2011, foram armazenados mais de 930 mil exames de imagens e de análises clinicas para o banco de dados da RCTM, que está em constante evolução. Ao todo são mais de 3700 usuários (médicos, enfermeiros, técnicos) que utilizam o sistema, sem considerar os acessos pelos pacientes (são 370 mil pacientes cadastrados no sistema que podem ter acesso aos seus exames). A Figura 2 apresenta a tela principal do sistema, onde podem ser visualizados todos os serviços disponíveis pela RCTM. Maiores detalhes sobre o funcionamento do sistema pode ser visualizado no Apêndice B.

Em função da grande quantidade de acessos simultâneos, o sistema pode ficar lento em horários de pico e por isso o envio e a recuperação da informação pode se tornar uma tarefa muito demorada. O modelo desenvolvido utiliza as técnicas tradicionais de Recuperação de Informação, baseadas em modelos Booleanos, e que não são eficazes na recuperação de documentos baseados em sentenças complexas

(Manning *et al.*, 2008) e consequentemente há uma necessidade cada vez maior de processar essas informações. Esse modelo exige também a aquisição de equipamentos cada vez mais complexos para poder comportar tal crescimento, como *Storages* de alta capacidade de armazenamento e sistemas de redes de alta velocidade.

Para solucionar esse problema de acesso ao grande número de informação, que se encontra disponível em diferentes bancos de dados, foi proposto neste trabalho o desenvolvimento de um modelo que utiliza técnicas de recuperação da informação mais inteligentes, com foco em conteúdo de informação semântica (Díaz-Galiano *et al.*, 2009). O objetivo aqui é ampliar o universo de pesquisa e possibilitar o reaproveitamento do conhecimento adquirido em atendimentos anteriores, de forma a facilitar o atendimento futuro.

1.2 DEFINIÇÃO DO PROBLEMA

Algumas tarefas comuns de investigação médica e de cuidados clínicos envolvem, por exemplo, a verificação dos resultados do exame, a comparação com outros relatórios ou análises estatísticas em busca de padrões de doença em prontuários de pacientes. Em situações como essas, onde é necessário que um grande volume de informação seja recuperado e analisado, torna-se crucial a automatização do processo para que se obtenha um resultado mais eficiente.

Conceitos médicos necessitam ser facilmente extraídos dos registros médicos e podem ser comparados com outras informações a fim de permitir buscas mais eficazes (Gschwandtner *et al.*, 2010). Embora a informação esteja disponível a partir de diferentes formas, os textos médicos precisam ser interpretados por computadores, de modo que a informação possa ser processada e efetivamente compartilhada. Para possibilitar esse processo, os usuários normalmente fazem uso de ferramentas computacionais a fim de aumentar a acessibilidade e o gerenciamento dos dados (Moskovitch e Shahar, 2009).

Por outro lado, interpretar o texto médico é uma tarefa difícil, mas pode ser mais fácil em comparação com o discurso narrativo, isso porque o vocabulário médico é mais restrito. Os registros clínicos médicos contêm um grande número de informações, normalmente escritos em texto livre e sem um padrão lingüístico. Médicos, em geral, não usam elementos de estilo para escrever seus laudos. Um médico pode escrever um laudo utilizando várias maneiras, cada profissional

possui seu próprio estilo de escrita e manipular essas informações é certamente um dos maiores desafios para os modernos sistemas de buscas na área de saúde (Sager *et al.*, 1987).

Uma vez que a informação em textos de laudos médicos normalmente não possui um padrão específico, o processo de recuperação de informações em documentos médicos é na maioria dos casos ineficiente. As tradicionais ferramentas de busca, por utilizarem técnicas Booleanas, não exploram todo o potencial existente nesse específico domínio de conhecimento (Moskovitch e Shahar, 2009). Para que toda a informação relacionada seja recuperada, é necessário que os mecanismos de busca sejam capazes de compreender a solicitação do usuário e expandir o universo de pesquisa, sem, no entanto perder a qualidade na pesquisa.

Por exemplo, considerando que um usuário que está utilizando um sistema hospitalar. Esse usuário pretende pesquisar na base de dados um caso de diagnóstico médico provido por outro profissional. Para efetuar essa consulta o usuário utiliza ferramentas computacionais para recuperar uma informação. As ferramentas de buscas normalmente recuperam somente as informações que foram digitadas pelo usuário no campo de pesquisa e eventuais erros de digitação têm um impacto direto nos resultados obtidos, o que acaba restringindo muito o universo de pesquisa.

Esse modelo também não possibilita a distinção de expressões sinônimas e nem de expressões negativas. Ou seja, quando o usuário solicitar à ferramenta de busca todos os documentos que contenham a expressão *"hipertensão arterial"*, o sistema retornará exclusivamente os documentos que contenham essas expressões. O sistema ignora expressões negativas existentes na base de dados. Documentos que contenham informações como *"o paciente não possui hipertensão arterial"* são retornadas como verdadeiras para o sistema de busca, mas para o usuário essa informação pode não ser relevante. Outro problema observado está relacionado aos documentos que contenham expressões como *"O paciente possui pressão arterial alta"*. Sentenças que não foram descritas explicitamente pelo usuário do sistema são excluídas da busca e consequentemente, não são apresentadas ao usuário.

Com o objetivo de solucionar esse problema, o trabalho aqui descrito visa desenvolver um modelo de recuperação de informações em bases de dados médicos, que utiliza técnicas de extração do conhecimento para anotar esses textos a partir de uma ontologia médica, detectar expressões negativas e expandir a pesquisa do usuário. Essa metodologia descreverá como conectar as entidades nomeadas a partir

de descrições da ontologia médica DeCS (Decs, 2010) e como expandir a consulta do usuário a partir dessa ontologia. Além disso, será desenvolvido um método para a detecção de expressões negativas automaticamente dos relatórios médicos utilizando técnicas de Processamento de Linguagem Natural (PLN), do inglês NLP (*Natural Language Processing*) (Gindl *et al.*, 2008). O uso das técnicas de anotação semântica, de expansão da consulta usando a ontologia médica e a detecção de expressões negativas em laudos e relatórios médicos, permitirá enriquecer semanticamente o tradicional sistema de Recuperação da Informação (do inglês, *Information Retrieval* – IR) e melhorar a qualidade da pesquisa do usuário.

A metodologia aqui proposta deverá ser capaz de responder às seguintes perguntas: mecanismos de buscas que utilizam conhecimento e semântica podem melhorar a precisão das respostas? Como recuperar a informação relevante à solicitação do usuário de forma mais precisa possível?

1.3 JUSTIFICATIVA E MOTIVAÇÃO

Para demonstrar um fluxo das atividades que um profissional médico pode enfrentar no seu dia-a-dia, a Figura 3 apresenta uma visão esquemática dessas tarefas. Para esse médico é possível que ele possa, por exemplo, emitir laudos de um exame, conferir laudos efetuados por outro profissional, ou emitir segunda opinião para um determinado laudo. Existe ainda, a possibilidade de ele consultar em diversas bases de literaturas médicas ou até mesmo na internet, para descobrir novas técnicas de investigação de procedimentos radiológicos/diagnósticos e terapêuticos ou até mesmo buscar novas técnicas de tratamentos de pacientes. Ainda, um médico pode apresentar palestras virtuais diretamente na *web* para uma equipe de atenção básica e interagir com esses profissionais a fim de prover educação continuada.

Tendo em vista a grande quantidade de tarefas que o profissional da saúde assume no seu dia-a-dia, a existência de uma ferramenta que o auxilie e agilize o acesso à informação é crucial para uma melhora na qualidade do serviço ofertado. Mas o armazenamento e a disseminação dessa informação não é uma tarefa simples, uma vez que processar grandes quantidades de informações exige o desenvolvimento de ferramentas de buscas com alto grau de precisão. E quando um grande volume de informação é capturado e analisado, a automatização dessas

tarefas é crucial para o processamento da informação de forma mais eficiente. Conceitos médicos devem ser facilmente extraídos dos registros médicos e esses conceitos podem ser comparados com outras informações a fim de permitir buscas mais eficazes (Gschwandtner *et al.*, 2010).

Embora a informação esteja disponível a partir de diferentes formas ou bases de dados, os textos médicos precisam ser interpretados por computadores, de modo que a informação possa ser processada e efetivamente compartilhada. Para permitir esse processo, os usuários precisam fazer uso de ferramentas computacionais para aumentar a acessibilidade e o gerenciamento de dados (Moskovitch e Shahar, 2009).

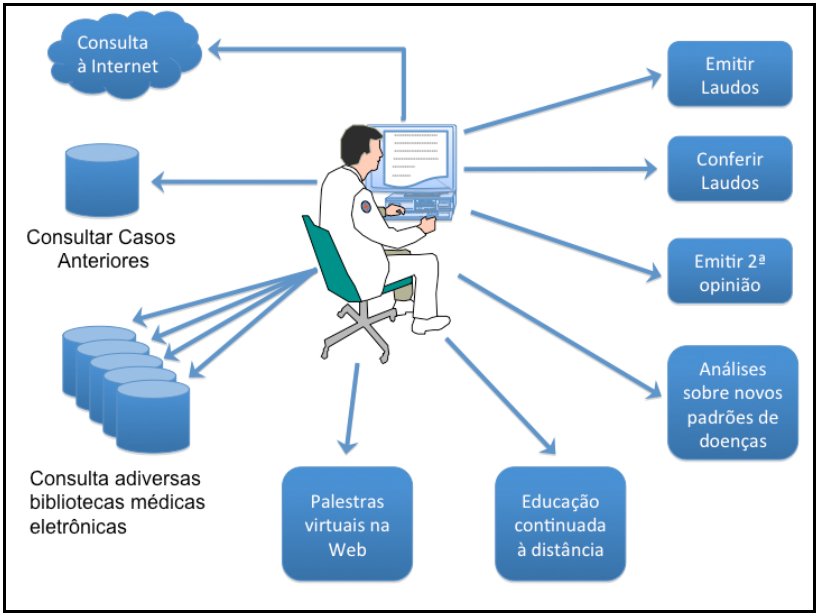


Figura 3: Tarefas comuns de um profissional de saúde no STT.

Problemas de armazenamento e acesso não são os únicos desafios para a utilização dos sistemas de telemedicina por parte dos profissionais na área médica. Existem ainda outras questões tecnológicas que impedem a sua utilização. Vários desafios devem ser superados para interpretar o conhecimento em prontuários médicos. A área de RI é apenas um deles e é o foco desse trabalho científico. Muitos trabalhos gerais de RI em documentos médicos não são adequados para prover qualidade na recuperação de informação (Hristidis *et al.*, 2007).

Segundo Hristidis et al., (2007) manipular declarações negativas em laudos médicos é um grande problema para as ferramentas de IR. Um resultado negativo em um diagnóstico médico é tão importante quanto um resultado positivo para a que o profissional médico tenha precisão em sua decisão. É muito comum em documentos médicos um diagnóstico negativo. Por exemplo, *"Ausência de nódulos no estudo da glândula da tireóide"*. Essa frase negativa gera um problema para os motores de buscas e a grande maioria dos buscadores exclui esses resultados da pesquisa, fazendo com que a resposta seja pouco confiável.

Partindo-se do pressuposto que essa lacuna ainda não fora preenchida no modelo utilizado pela RCTM, será desenvolvido um novo modelo de aquisição e comunicação do conhecimento médico, com o objetivo de melhorar o atual modelo de pesquisa, que é na maioria das vezes ineficaz na obtenção das respostas. Além disso, o atual modelo de pesquisa não permite recuperar a informação dos laudos dos exames.

Acredita-se que com o desenvolvimento de um sistema de pesquisa semântica poderá facilitar a interpretação dessas informações negativas e semi estruturadas das bases de laudos e prover maior precisão na resposta para o usuário. Pode-se entender que um sistema de recuperação de informação que possua uma base semântica, poderá contribuir para o preenchimento dessa lacuna no campo da medicina e permitirá que mais profissionais possam utilizar as tecnologias de telemedicina.

1.4 PRESSUPOSTOS OU HIPÓTESE DE TRABALHO

Partindo-se da premissa de que uma indexação de caráter semântico dos documentos, principalmente de laudo de um sistema de Registro Eletrônico de Saúde/Prontuário Eletrônico de Paciente/Portal de Telemedicina, possa agilizar e facilitar o acesso a informações, a presente pesquisa tem por objetivo desenvolver uma nova metodologia de busca em documentos médicos. Entende-se ainda, que as seguintes questões possam fortalecer as hipóteses dessa pesquisa:

- Há como utilizar ontologias médicas já existentes e aceitas pela comunidade médico-científica como indexadores para laudos (Dumas *et al.*, 2007)?

- É possível e eficiente utilizar-se da web semântica como forma de consulta inteligente às bases de dados indexadas pelas ontologias médicas (Berners-Lee *et al.*, 2001)?
- Sistemas de Recuperação de Informação podem melhorar o processo de busca por informação de forma a auxiliar o diagnóstico médico (Lourenço *et al.*, 2010)?
- Técnicas de expansão de pesquisas por detecção de frases negativas podem enriquecer o universo da pesquisa em bases de conhecimento médico (Díaz-Galiano *et al.*, 2009), (Gschwandtner *et al.*, 2010)?

Para melhor ilustrar as questões-chave deste trabalho será apresentada uma situação hipotética que ilustrará o funcionamento de um sistema baseado na metodologia a ser aqui desenvolvida.

1.4.1Cenário I – Recuperar informações do prontuário do paciente

Considere um hospital de grande porte que utiliza o armazenamento de documentos médicos textuais há mais de 10 anos e que atende em média 17.000 pacientes por mês. Esse hospital acumularia nesse período um pouco mais de um milhão de registros (levando-se em conta uma evolução histórica crescente de atendimentos)[1].

Segundo o Conselho Federal de Medicina – CFM é recomendado que as instituições armazenem todos os documentos dos pacientes por no mínimo 20 anos. Nesse caso, esse hospital, em 20 anos, deverá possuir mais de três milhões de registros armazenados em sua base de dados. As informações que estão armazenadas na base de dados servem normalmente para levantamentos estatísticos, ou para algumas pesquisas administrativas, ou então para pesquisas de algumas informações sobre dados de pacientes, como, nome, endereço, nome da mãe, etc.

Mesmo que os dados que estão armazenados na base estejam em formato de texto, eles são de pouca utilidade para os profissionais médicos quando buscam casos atendidos anteriormente, semelhantes aos que estão sob seus cuidados. Isso se dá em virtude da grande dificuldade e da demora em efetuar consultas nas atuais ferramentas de pesquisas

[1]Dados do setor de estatística do Hospital Universitário Professor Polydoro Ernani de São ThiagoHU–UFSC, para o ano de 2009.

desenvolvidas, visto que exige uma pesquisa sequencial em toda a base de dados.

Uma ferramenta que permita estruturar essas informações de modo a melhorar o processo de busca de diagnósticos anteriores pode ser fundamental para embasar ou compreender decisões de um novo caso. Mas a ausência dessa ferramenta desencoraja a utilização e o reaproveitamento do conhecimento adquirido que fora armazenado nas bases de dados do sistema legado. Essas informações poderiam ser utilizadas, na maioria dos casos, para dar maior agilidade ao tratamento e a recuperação dos pacientes. Além disso, de posse dessa ferramenta o profissional pode evitar tratamentos anteriores ineficientes ou indicar novas formas de tratamentos, além de possibilitar uma grande redução de custos para o hospital.

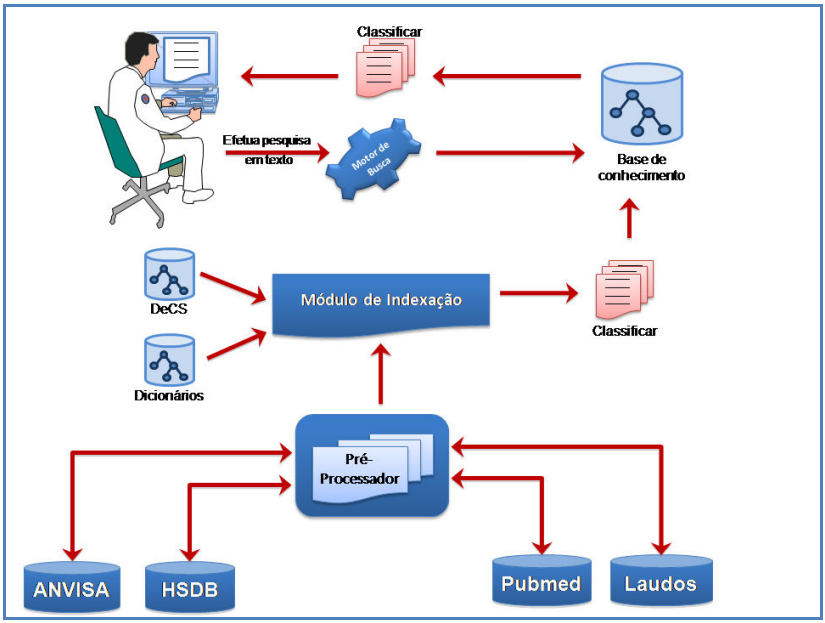


Figura 4: Modelo proposto.

O novo modelo de pesquisa, apresentado por meio da Figura 4, permitirá que um usuário médico ou um profissional de saúde possa recuperar as informações de uma base de conhecimento de forma mais rápida que a atual tecnologia. Para a criação dessa base de conhecimento, será desenvolvido um módulo chamado "pré-processador" que irá recuperar as informações da base de laudos do

Sistema Catarinense de Telemedicina e Telessaúde, normalizar e enviar para o módulo de indexação. Esse módulo é responsável por expandir a pesquisa do usuário, detectar expressões negativas e anotar as informações de doenças utilizando a ontologia DeCS e dicionários semânticos e léxicos.

A partir da indexação, o conhecimento adquirido será classificado e armazenado na base de conhecimento (Cabral, 2010). Com o uso dessa metodologia, um profissional poderá efetuar pesquisas dentro dos laudos dos pacientes, buscar um determinado diagnóstico efetuado no passado, que poderá ajudar no tratamento futuro de um paciente. Ainda, esse modelo proposto possibilitará a recuperação de informações sobre um termo solicitado seus termos sinônimos e termos relacionados, além de permitir a comunicação do conhecimento que fora armazenado no sistema, para uso em pesquisas futuras.

1.4.2 Cenário II – recuperar informações de atendimentos realizados anteriormente

Considere um sistema que permita a promoção do desenvolvimento contínuo dos profissionais das Equipes de Saúde da Família (ESF) a distância com a utilização de multimeios, como segunda opinião formativa, palestras virtuais, ensino a distância e apoio a essas ESFs. Nesse modelo, conforme apresentado na Figura 5, podem participar médicos da família, enfermeiros, dentistas, psicólogos, assistentes sociais, fisioterapeutas, farmacêuticos e nutricionistas para dar palestras via *web*, promover a capacitação continuada e o atendimento de segunda opinião formativa às equipes da saúde da família.

As equipes dos municípios credenciados recebem apoio remoto dos serviços por meio de um aplicativo chamado Painel de Discussões (tele consultorias e suporte a dúvidas clínicas) e do ensino à distancia - EAD (capacitações, disponibilização de material de aprendizagem multimídia e de alto rigor científico). Esse painel de discussões é representado na Figura 5 pelo *"Núcleo de Telessaúde HU/UFSC"* e promove a integração entre profissionais médicos da Atenção básica com os especialistas de domínio.

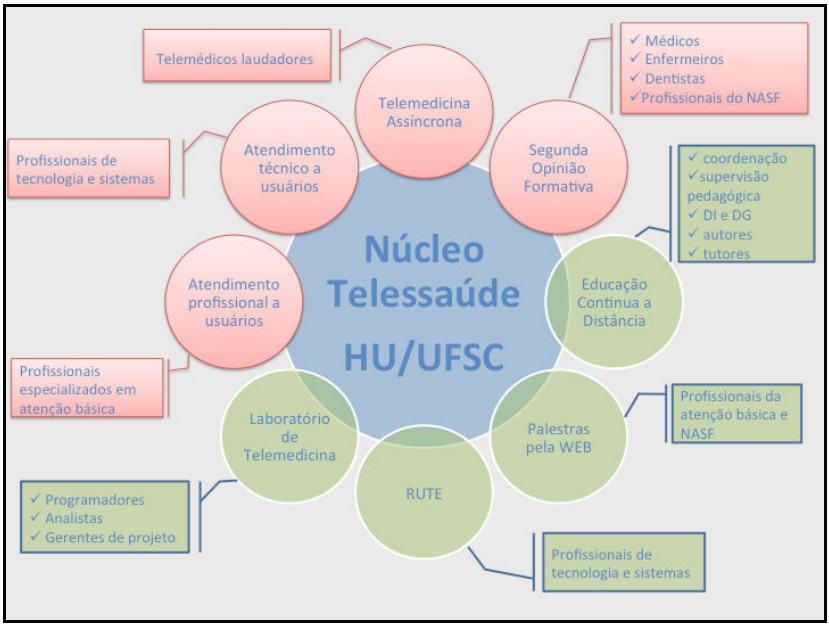


Figura 5: Integração do sistema de Telemedicina e Telessaúde de SC

.

Normalmente, as situações-problemas do dia-a-dia podem ser discutidas com profissionais da atenção básica que estão conectados ao sistema e, quando necessário, um especialista poderá efetuar a complementação de uma discussão de casos por meio da segunda opinião e o profissional de atenção básica pode utilizar o tele-diagnóstico por imagem nas Unidades Básicas de Saúde para ampliar seus conhecimentos sobre o assunto.

Todas as discussões trocadas entre os profissionais são armazenadas no banco de dados do painel de discussões, que podem ser utilizadas em consultas futuras. Por exemplo, quando um usuário necessita de uma opinião sobre o uso racional de medicamentos. Esse usuário pode efetuar sua pergunta por meio da interação com o Centro de Informações Toxicológicas que faz parte do sistema, ou procurar pela informação diretamente na base de dados de atendimentos em toxicologia. Nesse segundo caso, a ferramenta de busca deve ser capaz de retornar a informação desejada para o usuário da forma mais adequada possível e permitir efetuar a busca não somente na base de toxicologia, mas em todo o sistema.

Como resultado, o usuário não precisará mais acessar diversas bases para encontrar a informação desejada. O desenvolvimento dessa

ferramenta possibilitará a recuperação, o reuso do conhecimento e ainda melhorar o desempenho das consultas de forma totalmente transparente para o usuário.

1.4.3 Cenário III – Estatísticas de morbidade

Considere que o setor de vigilância epidemiológica do Estado de Santa Catarina deseja saber sobre a quantidade de portadores de uma doença em relação à população de determinada região. Por exemplo, um usuário do sistema deseja saber qual a quantidade de pacientes que contenham *"Insuficiência Cardíaca"* nos laudos de seus exames. Ao efetuar uma busca por esse termo, a ferramenta de busca do sistema pesquisará não somente por esse termo, mas também pelos termos como, *"Falência Cardíaca"*, *"Descompensação Cardíaca"* e *"Falência Cardíaca Congestiva"*. Essas informações são sinônimas ao termo solicitado pelo usuário e são encontradas na base de conhecimento do sistema. Esse conhecimento adquirido se deve ao fato do relacionamento entre os termos encontrados na ontologia DeCS com os termos da base de dados.

Como resultado dessa pesquisa, o usuário saberá qual região do Estado apresenta maior ou menor número desses achados. De posse dessas informações é possível saber também qual o termo é mais utilizado para definir um achado ou saber se diferentes médicos definem achados com expressões diferentes (sinônimas) em cada região do Estado. Ainda, o sistema permite que laudos que contenham expressões negativas possam ser excluídos ou incluídos na pesquisa do usuário e, dessa forma, tornando a precisão do sistema ainda maior. De posse desses dados, o usuário terá muito mais informação sobre a expressão pesquisada e consequentemente os resultados serão mais confiáveis.

1.5 OBJETIVOS

1.5.1 Objetivo Geral

Desenvolver um modelo diferenciado para recuperação e comunicação do conhecimento, a partir de bases de dados médicos, que permitirá ampliar a consulta do usuário e melhorar o processo de busca

em relatórios médicos, com o intuito de obter melhor precisão nas respostas.

1.5.2 Objetivos Específicos

Identificar pontos fortes e fracos nos modelos de RI empregados na área médica;

Propor um modelo de pesquisa para a recuperação e comunicação do conhecimento, a partir de diferentes bases de dados de forma a aperfeiçoar o tradicional processo de busca e indexação;

Demonstrar a viabilidade do modelo proposto por meio da elaboração de um protótipo, assim como a aplicação desse modelo em um estudo de caso;

Definir normas para a criação de um conjunto de procedimentos para avaliação e testes do protótipo desenvolvido;

Realizar uma análise comparativa do modelo proposto contra o modelo tradicional de busca.

1.6 ESCOPO E DELIMITAÇÃO DO TRABALHO

Para que seja realizável a recuperação de dados médicos de maneira unificada, faz-se necessário, como descrito anteriormente, o emprego de diversas técnicas para normalização desses dados. O emprego de tecnologias para representação e recuperação do conhecimento, fundamentadas em ontologias de termos médicos, pode fornecer o meio para obtenção de tal objetivo. Dessa forma, é possível fundamentar a base metodológica a ser seguida no decurso deste trabalho.

Como primeira etapa, propõe-se executar um levantamento das diferentes formas de buscas em dados médicos disponíveis a titulo de material para este trabalho. Faz parte do escopo o estudo de técnicas para expandir as pesquisas tradicionais usando como base a ontologia médica DeCS. Não faz parte do escopo desse trabalho o estudo de outras ontologias médicas, como SNOMED, ou UMLS, por entender que essas ontologias não fazem parte da rotina diária da equipe médica brasileira.

Serão estudadas as técnicas de anotações semânticas em bases de conhecimento usando entidades nomeadas para definição dos termos que serão utilizados nos documentos médicos. Em um primeiro

momento, as anotações serão feitas de forma automática utilizando o DeCS como referencial. A partir desse levantamento das entidades nomeadas encontradas na base, será utilizado um especialista médico para validar os termos encontrados nos documentos médicos. Para os termos que não puderam ser extraídos do DeCS, será criado um dicionário léxico para o reconhecimento dessas entidades nomeadas que serão correlacionadas aos termos da base de conhecimento (aqui essa base será chamada de conhecimento cotidiano). Não serão estudadas as técnicas de construção de ontologias para a definição desse dicionário, por entender que as ontologias desenvolvidas na área médica já estão extremamente maduras e livres de ambigüidades.

Faz parte do escopo dessa pesquisa a identificação de expressões negativas utilizando um algoritmo baseado em expressões regulares para determinar se achados ou doenças mencionados em relatórios médicos estão negados ou não. Não faz parte do escopo desse trabalho o estudo das diversas técnicas de NLP utilizadas para interpretar semanticamente os textos médicos.

Também não faz parte do escopo deste trabalho o estudo dos procedimentos legais para acesso às informações, a análise das regras de acesso aos dados de pacientes, nem o desenvolvimento de técnicas para a definição de novas regras para estruturar uma base de dados médico. O estudo aqui apresentado trata da recuperação da informação a partir da base de dados do STT e da base de toxicologia clínica do CIT/SC, que já se encontram sob sigilo médico e estão amparados pelas regras médicas e pelas regras da segurança da informação.

Do ponto de vista da utilização de ontologias, será desenvolvida uma metodologia para executar a normalização e indexar as informações dos dados de pacientes. Entende-se por normalização de dados a representação da informação sob a forma de algum padrão (Clunie, 2000), que permita acessar de maneira determinística, sem a necessidade de utilização de conversores externos, e que seja passível de ser automatizado pela tecnologia desenvolvida pelo Grupo Cyclops (Wallauer *et al.*, 2008).

Recuperar dados de diagnósticos de uma forma ergonômica e rápida também faz parte do escopo deste trabalho. Para executar a recuperação, acredita-se, com base em estudos preliminares (Berners-Lee *et al.*, 2001), que a utilização da web semântica forneça a estrutura necessária para compor tanto a base ergonômica, quanto suprir os requisitos de velocidades no que se refere à recuperação das informações. Nessa etapa será utilizada a infra-estrutura do STT como base de testes. Não faz parte do escopo deste trabalho o estudo de

técnicas de acesso à banco de dados para a comunicação entre clientes e servidores, nem implementações da web semântica clássica, como definição de ontologias, OWL, RDF ou utilização de linguagem XML.

Para a validação da proposta, pretende-se utilizar as métricas de avaliação de sistemas de RI (*precision* e *recall*) no sentido de avaliar os resultados obtidos e compará-los com as tradicionais ferramentas de buscas desenvolvidas na área. Duas das principais métricas serão utilizadas: a *Average Precicion,* que representa uma média de precisão calculada sobre um conjunto de consultas (Buckley e Voorhees, 2000) e a *Precision at Ten (P@10)*, que considera os dez primeiros registros para avaliar a precisão (Van Rijsbergen, 1975).

1.7 RESULTADOS ESPERADOS

O resultado esperado dessa pesquisa será a criação de um modelo de pesquisa e comunicação do conhecimento médico, incluindo-se bases de laudos de pacientes, informações toxicológicas e bases de artigos científicos disponível em bibliotecas digitais.

O primeiro resultado desse trabalho refere-se à criação de uma base de conhecimento médica baseada na ontologia médica DeCS que será utilizada como arquivo de índice para novas buscas e também para a comunicação e a disseminação do conhecimento médico adquirido. Nessa base de conhecimento estarão inseridos os ativos de conhecimentos que foram extraídos do método de anotação semântica e que referenciaram as entidades nomeadas da ontologia DeCS com os termos encontrados na base de laudos do STT.

O segundo resultado esperado possibilitará a criação de um modelo que, utilizando como base a expressão descrita pelo usuário, permitirá expandir o universo de pesquisa desse usuário adicionando os termos sinônimos e os itens relacionados da ontologia DeCS à pesquisa inicial de forma totalmente transparente para esse usuário.

O terceiro resultado esperado será o desenvolvimento de um modelo de detecção de expressões negativas e a composição de um dicionário com termos e expressões negativas mais utilizadas em vocabulários médicos. Esse modelo permitirá que termos e frases negativas possam, ou não, compor a pesquisa do usuário. A pretensão aqui é a ampliação do universo de pesquisa na base de conhecimento e um aumento na precisão das respostas ao usuário.

Como último resultado para o desenvolvimento desse trabalho será apresentada uma análise comparativa da nova técnica de pesquisa e recuperação da informação contra outros modelos relacionados. Ainda, serão produzidos artigos científicos para comprovar a teoria desse novo modelo de recuperação de informações.

1.8 INEDITISMO E CONTRIBUIÇÃO CIENTÍFICA

Atualmente tem-se desenvolvido diversas técnicas para a extração de informações em bases de conhecimento, mas durante o período de estudo dos trabalhos relacionados, não foram encontradas pesquisas com os mesmos objetivos desse trabalho. Cada pesquisa é focada somente em uma área pré-estabelecida. Normalmente as pesquisas seguem na área e expansão de pesquisas, mas não levam em consideração expressões negativas para aumentar a precisão de uma busca. Técnicas de buscas que estudam expressões negativas não levam em consideração as buscas semânticas e as técnicas de pesquisas baseadas em anotações semânticas não utilizam as buscas expandidas para aumentar a precisão de suas buscas.

A contribuição cientifica inédita que esse trabalho visa chegar consiste na concepção de um modelo diferenciado para a recuperação e para a comunicação do conhecimento a partir de bases de dados médicos, que permitirá ampliar a consulta do usuário e melhorar o processo de busca em relatórios médicos, com o fim de se obter melhor precisão nas respostas para o usuário.

Esse novo modelo será integrado ao STT, que contribuirá para o desenvolvimento de um portal integrando em uma única plataforma on-line a coleta dos dados ou imagens de um exame, a remessa dos resultados, a solicitação de internação e as decisões a serem tomadas por profissionais de saúde, além de uma ferramenta única de pesquisa e recuperação de informações. É uma excelente maneira de unificar todas as funcionalidades da Telemedicina Assistencial.

A indexação e consulta semântica de laudos médicos, prática hoje ainda não explorada, será implementada de forma mais rápida que o atual modelo de busca tradicional. Dessa maneira, a utilização de tecnologias da web semântica (Berners-Lee *et al.*, 2001), juntamente com as técnicas de RI (Greengrass, 2001) e a expansão de pesquisas utilizando ontologias médicas (Díaz-Galiano *et al.*, 2009) contribuirão de forma muito significativa o desenvolvimento deste trabalho.

Nesse modelo será definida uma nova teoria correlacionando a linguagem médica utilizada em laudos, um sistema de busca inteligente e uma base de conhecimento médico que será utilizado em consultas semânticas. Será criado um novo método de fazer consultas, pois o usuário do sistema será capaz de fazer buscas com resultados muito mais completos e não somente buscas por palavras-chave.

O modelo visa oferecer uma técnica de detecção de frases negativas utilizando regras da NLP e permitirá também a criação de um dicionário médico com as expressões negativas mais utilizadas pela comunidade médica, baseado nos textos contidos na base de dados do STT e validados por especialistas do domínio médico.

Outra contribuição refere-se ao uso de uma nova metodologia que permitiu expandir as pesquisas médicas, relacionando termos sinônimos e conceitos relacionados que são extraídos da ontologia médica DeCS (Díaz-Galiano *et al.*, 2009).

1.9 CARACTERIZAÇÃO DA MULTIDISCIPLINARIDADE

A presente pesquisa está inserida na área de Engenharia e Gestão do Conhecimento (EGC) e apresenta um modelo para a recuperação e comunicação do conhecimento em documentos médico. O caráter multidisciplinar desta pesquisa é dado por meio da ação compartilhada das técnicas da Engenharia do Conhecimento (EC), da Gestão do conhecimento (GC) e da computação para aperfeiçoar o conhecimento em saúde dos profissionais médicos.

Uma das principais características da GC é definida por um conjunto de processos que governam a criação, o uso e a disseminação de conhecimento em uma organização com o objetivo de atingir suas metas.

A GC possui ainda o objetivo de controlar, facilitar o acesso e manter um gerenciamento integrado sobre as informações em diversos meios. O conhecimento é definido como a informação interpretada, ou seja, o significado de cada informação pode ser utilizado para importantes ações e para tomadas de decisões.

Com o objetivo de auxiliar no gerenciamento das informações, os Sistemas de Conhecimento (SC) são propostos para apoiar os processos de criação, armazenamento, recuperação comunicação e aplicação de conhecimento. Esses sistemas são baseados em Tecnologia de

informação e por meio dessas técnicas, a EC visa desenvolver sistemas de conhecimento que permitam apoiar a GC em uma organização.

O programa de Pós-Graduação em Engenharia e Gestão do Conhecimento (PPGEGC) visa: *"a pesquisa de novos modelos, métodos e técnicas de engenharia, de gestão e de disseminação do conhecimento para as novas organizações e para a Sociedade."* (Egc, 2011). Essas três áreas do conhecimento devem trabalhar em conjunto para promoverem o compartilhamento do conhecimento entre si e um sistema inteligente é desenvolvido utilizando-se técnicas de EG, com a ajuda de especialistas de domínio do conhecimento. A EG é responsável pela extração e representação do conhecimento por meio da utilização de linguagens de IA (Studer *et al.*, 1998).

A presente tese está inserida na área da EG e apresenta um modelo que utiliza metodologias computacionais bem fundamentadas teoricamente e ainda outras atualmente idealizadas, porém ainda não propostas, com o objetivo de promover a representação, a recuperação e a comunicação do conhecimento em dados médicos.

Esse envolvimento multidisciplinar entre a medicina e a informática, além de ser imprescindível à execução do trabalho, dará oportunidade à equipe de pesquisadores de ampliar a visão da realidade e do potencial do tema em relação ao projeto proposto.

Dessa forma, a multidisciplinaridade fica caracterizada como uma integração entre as facilidades computacionais para manipulação de dados e a necessidade médica de gerenciamento de sua base de informação.

1.10 ESTRUTURA DO TRABALHO

Conforme apresentado anteriormente, esse trabalho de pesquisa objetiva o desenvolvimento de um modelo para a recuperação e comunicação do conhecimento em documentos médicos. E para chegar aos resultados deste trabalho foram definidos diversos procedimentos metodológicos que serão apresentados conforme a organização dos capítulos dessa tese.

Primeiramente, no capitulo 1 foi apresentado a introdução desse trabalho, onde são tratados assuntos como a definição do problema de pesquisa, justificativa, motivação, objetivos e resultados esperados. Além disso, para melhor entendimento do leitor são apresentados três cenários de aplicação do modelo aqui estudado.

A seguir, no capítulo 2, o trabalho inicial constituiu em estudar as técnicas para aquisição e representação do conhecimento em saúde. Nesse capítulo são discutidos como as ontologias são estruturadas e como elas podem ajudar a enriquecer o conhecimento médico. Como resultado desse levantamento, chegou-se a conclusão que ontologias médicas podem ser usadas em sistemas de conhecimento como direcionador em ferramentas de pesquisas em bases de dados médicas. Nesse sentido, foram pesquisados quais os modelos e técnicas de recuperação de informação são utilizadas para auxiliar o processo de busca médico. Os estudos sobre essas técnicas de RI são apresentadas no capítulo 3.

Em seguida, no capítulo 4, foi realizada uma revisão bibliográfica sobre os problemas e soluções propostas a fim de entender o que está sendo pesquisadas no mundo na área de expansão de pesquisa, recuperação de frases negativas e anotação semântica dentro da área médica.

O capítulo 5 apresenta o modelo proposto. Nesse capítulo é discutida primeiramente uma visão alto nível da presente proposta, seguido da organização do modelo conceitual do sistema e dos componentes que foram utilizados para a construção do protótipo. Ainda, o funcionamento do modelo proposto e como o conhecimento é adquirido e representado, são descritos em detalhes no capitulo 5. A descrição da implementação do protótipo é dada nesse mesmo capítulo.

Para avaliar a funcionalidade do modelo, um conjunto de regras foi definido no capítulo 6, onde vários laudos médicos foram validados por especialistas a fim de criar uma base de conhecimento anotada. Os testes experimentais e os estudos realizados, bem como os resultados obtidos são discutidos no capítulo 7 e, finalmente, as conclusões e trabalhos futuros são apresentadas no capitulo 8.

2 AQUISIÇÃO E REPRESENTAÇÃO DO CONHECIMENTO EM SAÚDE

O uso e a disseminação do conhecimento são caracterizados por um conjunto de processos que tem a finalidade de auxiliar as organizações a gerenciarem seus conhecimentos. Tais processos são chamados de Gestão do Conhecimento (GC), definidos como sistemas baseados em Tecnologia de Informação (TI), desenvolvidos para embasar os processos de criação, armazenamento, recuperação, comunicação e aplicação de conhecimentos. Nesse sentido, a TI pode ser considerada como um conjunto de atividades e soluções providas por recursos computacionais com o objetivo de difundir o conhecimento (Manica, 2009).

A TI permite acesso a serviços independentemente da localização geográfica e da condição social dos indivíduos que a utilizam. Permite divulgar pesquisas científicas, melhorar a qualidade e a disponibilização das informações, agregar valor aos serviços e produtos ofertados. Na área médica, pode-se perceber que a TI está fortemente ligada à evolução das pesquisas científicas, visto que a prestação de cuidados à saúde é um processo que exige intenso conhecimento relativo aos pacientes, diagnósticos, tratamentos e outros fatores que influenciam na tomada de decisão ou na gestão de recursos em saúde (Manica, 2009).

Dentro dessa perspectiva, a EC propõe o uso de modelos computacionais que possibilitam descrever o raciocínio humano em áreas específicas. Na área da saúde, os chamados Sistemas Baseados em Conhecimentos (SBC), ou Sistemas Especialistas (SE), proporcionam ao computador "entender" o conhecimento médico especializado e apoiar a gestão do conhecimento em uma organização. O conhecimento tem como foco principal as atividades profissionais e os procedimentos decisórios. Ele está inserido na maioria das tarefas executadas por profissionais em saúde e dificilmente ocorre de forma isolada. Esse conhecimento pode ser gerado durante a prática das atividades médicas, adquirido a partir de fontes diversas (jornais, revistas, periódicos), ou até mesmo em conversas informais (Landry *et al.*, 2006).

A capacidade de adquirir, criar, compartilhar e aplicar o conhecimento é essencial para resolver problemas em saúde pública. O conhecimento é resultado de uma série de transformações que vão desde o armazenamento dos dados sobre uma realidade, até a interpretação das informações a fim de se obter uma ação (Landry *et al.*, 2006). A capacidade de adquirir o conhecimento consiste na extração do

conhecimento de um especialista, ou a partir de bibliografias confiáveis e transpor para um sistema computacional com a finalidade de torná-lo inteligente. Mas esse processo pode ser dispendioso, pois muitas vezes o conhecimento não está expresso em livros ou em manuais (conhecimento explícito). Ele pode estar contido nas experiências, emoções e ações dos profissionais médicos. Esse conhecimento é chamado de conhecimento tácito. Segundo Nonaka e Takeuchi (1997) o conhecimento tácito é na maioria das vezes inexpressível difícil de formalizar e isso dificulta a transmissão e o compartilhamento. Por exemplo, quando um profissional médico reconhece uma série de sintomas de um paciente, aparece em sua mente imediatamente um conjunto de diretrizes para determinar um diagnóstico e para escolher o melhor tratamento a ser ministrado. Esse é o princípio básico do raciocínio clinico e baseia-se em estudos de casos passados, tentativas e erros, e intuição para tratar o paciente (MOURADIAN, 1990).

A partir desse exemplo pode-se observar que a aquisição do conhecimento ainda apresenta dificuldades em representar o raciocínio médico. Nesse sentido, Durkin (1984) propõe o uso de diversas técnicas, como estudos de casos e entrevistas para resolver os problemas relacionados à imprecisão das informações nas áreas médicas. A representação do conhecimento é a parte essencial dos sistemas inteligentes. Ela resulta na expressão dos pensamentos, experiências, observações e metodologias aplicadas pelos especialistas do domínio. É nesse cenário que a EC atua. O objetivo da EC é prover técnicas e métodos para converter o conhecimento tácito em conhecimento explícito.

Um sistema de representação do conhecimento de especialistas humanos deve possuir um conjunto de informações para que possa ser capaz de resolver problemas de forma criativa, correta e eficaz. Um sistema inteligente precisa ter o conhecimento do contexto em estudo e saber reconhecer os processos de mudança dos fatos, para poder encontrar possíveis soluções, juntamente com algumas estratégias de como solucionar cada problema. O estudo da representação de conhecimento deve ser capaz de entender os problemas para poder codificá-los em um programa computacional (Kong *et al.*, 2008).

Para organizar e representar o conhecimento, (Manica *et al.*, 2009) sustenta que a EC apresenta diversas ferramentas terminológicas, que vão desde um simples dicionário controlado até uma sofisticada ontologia. Por esse motivo, a utilização de conhecimento sobre um determinado domínio previamente organizado pode representar ganhos muito mais significativos. A utilização dessas ferramentas pode fornecer

um vocabulário que, se estiver bem definido, estabelece um consenso terminológico a ser utilizado e define os conceitos referenciados. Manica (2009) afirma ainda que:

> Na área da saúde, as ontologias são normalmente utilizadas para auxiliar a troca de informações clinicas entre os sistemas computacionais e no desenvolvimento de novas aplicações como prontuário eletrônico, segunda opinião diagnóstica, sistemas de suporte e decisão clinica, dentre outros2.1 ONTOLOGIAS

Uma ontologia consiste em um conjunto de conceitos, relacionamentos entre eles e regras que regem estes relacionamentos. É uma forma de representar um conjunto de objetos e suas relações, possibilitando, assim, que um sistema computacional entenda não apenas a sintaxe desse conjunto de objetos, mas também sua semântica. Ontologias em sistemas de informação clínica são muito utilizadas como forma de representar conceitos médicos e os relacionamentos entre eles. Linguagens para descrição e métodos para a unificação de ontologias têm sido desenvolvidos com o objetivo de facilitar o compartilhamento de informações entre instituições médicas (Berners-Lee *et al.*, 2001).

Uma ontologia é basicamente constituída por classes, relações, axiomas e instâncias. As classes, também chamadas de conceitos, podem representar qualquer coisa em um domínio, como por exemplo, uma estratégia ou uma tarefa qualquer. As relações constituem uma forma de interação entre as classes no domínio. Axiomas podem representar as sentenças verdadeiras. Já as instâncias são utilizadas para compor os elementos do domínio (Dumas *et al.*, 2007). A classificação das ontologias pode ser executada de acordo com o grau de formalidade de seu vocabulário, conforme sua estrutura, assunto da conceitualização, função e aplicação. Uma ontologia sempre compreende um vocabulário de termos e a discriminação de seu significado.

As ontologias são capazes de representar fontes de dados, oferecendo uma maior organização e uma melhor recuperação. Uma compreensão comum e compartilhada de um domínio é possibilitada por uma ontologia, propiciando um compartilhamento do conhecimento das pessoas e com os sistemas.

Seguindo nessa linha, alguns autores consideram que a definição de Gruber (1993) sobre ontologia, como sendo uma especificação explícita de uma conceitualização expressa a partir de uma linguagem formal e tendo uma visão abstrata e simplificada do mundo que ora fora

representado. Os autores, ainda, classificam as ontologias em diversas variações, de acordo com o seu nível de expressividade. Essa classificação pode descrever um baixo nível semântico, até instrumentos que possam conter relações semânticas mais complexas.

Na literatura, há diversos artefatos que podem ser considerados ontologias, tais como vocabulários controlados, descritores, sistemas terminológicos, terminologias léxicas, entre outros (Lassila e Mcguinness, 2001). E, dentro do contexto da Ciência da Computação, é possível construir ontologias utilizando uma perspectiva dedutiva, ou seja, tendo como base uma ontologia mais genérica de um domínio é possível criar uma nova ontologia mais restrita e específica, dentro do mesmo domínio de conhecimento.

Em 2005, a organização norte-americana *National Information Standards Organization*, definiu regras para a construção, a formatação e a manutenção de vocabulários controlados monolíngües (Ansi/Nisoz39-19-2005, 2005). Esse documento define um vocabulário controlado como uma lista finita de termos que tem seus respectivos significados explícitos com o objetivo de evitar redundâncias e ambigüidades, e são utilizados para representar informações seguindo um padrão pré-estabelecido. Vocabulários controlados possuem estruturas para permitir que diferentes tipos de relacionamentos entre termos, possam ser determinados desde níveis de relacionamentos mais simples, até estruturas mais complexas. A Figura 6 apresenta os níveis de complexidade de um vocabulário controlado, que vão desde uma simples lista até um *thesaurus*, passando pela lista de sinônimos e taxonomias.

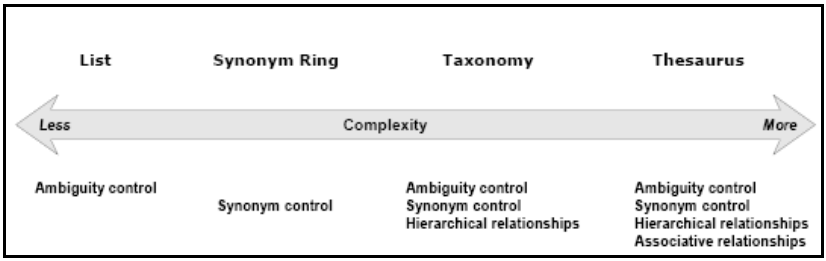


Figura 6: Níveis de complexidade de um vocabulário controlado.
Fonte: Ansi/Nisoz39-19-2005 (2005).

Conforme os autores afirmam, uma taxonomia pode ser definida como uma coleção de termos de um vocabulário controlado que está organizado em uma estrutura hierárquica, mas que não possibilita

atribuir características ou propriedades de tais termos, nem manifestar outros tipos de relacionamentos (Ansi/Nisoz39-19-2005, 2005).

Ainda segundo o relatório técnico ANSI/NISO Z39-19-2005 (2005), um *thesaurus* é um vocabulário controlado organizado com base em uma ordem conhecida e estruturado com o objetivo de disponibilizar claramente os relacionamentos de equivalência, associação, hierarquia e homônimos existentes entre os termos (por exemplo, o SNOMED, que será apresentado na seção 2.2.1 SNOMED). Um *thesaurus* pode conter características de taxonomias, como um conjunto de relacionamentos semânticos, que permitem que os conceitos e relacionamentos possam ser descritos de maneira consistente em uma classificação ou em um sistema de RI.

Lassila e McGuinnes (2001) apresentam um esquema classificatório das diversas variações que o termo ontologia pode assumir, baseado na estrutura e no conteúdo da ontologia. Essa classificação descreve desde uma ontologia simples, como por exemplo, um catálogo, até uma ontologia que contenha relações semânticas mais complexas. Mas todos os tipos de ontologias têm um único propósito: estabelecer um vocabulário compartilhado, com diferença em grau de formalismo e de expressividade de cada representação.

As ontologias apresentadas por Lassila e McGuinnes (2001) são descritas das mais simples (que requerem um nível de expressividade baixo) para as mais complexas (nível de expressividade alto), conforme segue:

- **Vocabulários controlados ou Catálogos**: uma lista finita de termos e seus respectivos significados que são utilizados para representar informações em um contexto específico;
- **Termos ou Glossários**: uma lista de termos e seus respectivos significados em linguagem natural, análogos a um dicionário;
- ***Thesaurus***: um vocabulário controlado que oferece também relacionamentos entre esses termos;
- **Hierarquias informais**: hierarquias que utilizam relacionamentos informais, de forma a permitir incorporar conceitos a uma determinada categoria, mesmo que formalmente esses conceitos não façam parte dessa categoria.
- **Hierarquias formais**: hierarquias que contêm instâncias de um determinado domínio, de forma a permitir que os relacionamentos sejam respeitados na essência e descritos a partir de taxonomias;
- **Frames**: ontologias que incluem classes e propriedades, de modo que as propriedades não possuam escopo global, sendo aplicados

somente nas classes para as quais elas foram definidas e possibilitam contextualizar as informações em um domínio específico;

- **Restrição de valores**: tipos de ontologias que definem restrições para os valores assumidos nas propriedades de suas classes;
- **Restrições lógicas**: tipo de ontologias que possibilitam a definição de restrições lógicas, com o objetivo de beneficiar a realização de inferências automatizadas.

A partir das hierarquias formais, (Lassila e Mcguinness, 2001) consideram essas definições como sendo ontologias semânticas mais complexas e para um artefato ser considerado uma ontologia é indispensável que contenham um vocabulário controlado finito de termos, interpretação não ambíguas de classes e relacionamentos entre os termos desse vocabulário e relacionamentos hierárquicos precisos entre as classes e subclasses.

A partir das definições apresentadas, esse trabalho de pesquisa considera as terminologias, os glossários, os *thesauri*, os vocabulários controlados e os descritores em saúde como sendo ontologias com um baixo nível semântico.

2.2 ONTOLOGIA MÉDICA

Existem diferentes técnicas que são utilizadas para representar e compartilhar o conhecimento dos especialistas de um determinado domínio. Dentre eles destacam-se dicionários léxicos, taxonomias, *thesaurus e* ontologias. Dessas ferramentas, as que mais são utilizadas atualmente para o compartilhamento do conhecimento entre profissionais de saúde são as ontologias. As ontologias são um importante meio de representar, formalizar e compartilhar conhecimento, para que possa ser reutilizado posteriormente por outras aplicações (Rubin *et al.*, 2008).

Na área da saúde, o uso de ontologias tem destaque especial para auxiliar a comunicação de informações entre sistemas computacionais e também no auxilio ao desenvolvimento de aplicações médicas, como sistemas de apoio a decisões, ou sistemas de telemedicina. Manica (2009) acrescenta que*:*

> As ontologias medicas são um recurso importante para o desenvolvimento da medicina baseada na evidencia, pois além de incorporarem dados de saúde, introduzem especificações formais para

representar relacionamentos estruturais entre os termos.

As seções seguintes apresentam as ontologias mais utilizadas para representar o conhecimento dentro da área de saúde.

2.2.1 SNOMED

SNOMED (*Systematized Nomenclature of Medicine*) é uma das mais completas nomenclaturas multiaxiais criadas para indexar o conjunto de registros médicos, possui tradução em diversos idiomas (Alemão, Espanhol e Inglês) e, em 2008, ela possuía mais de 311.000 conceitos com significados únicos e definições formais baseadas em hierarquias. Essa lista de nomes ou conceitos está organizada segundo tipos semânticos e hierárquicos de classes de objetos. A SNOMED internacional foi formada em setembro de 1993, mas já havia sido traçada desde o início dos anos 60, como a SNOP *(Systematized Nomenclature for Pathology)* (Snomed, 2010).

Um aspecto peculiar na ontologia SNOMED é que ela é composta por 19 eixos hierárquicos e várias subclassificações. A categorização é feita de acordo com a classe semântica que pertence determinado conceito. Está dividida em conceitos, em hierarquias, em relacionamentos e em descrições. Dentre as 19 hierarquias e sub-hierarquias, a SNOMED possui quase 1,45 milhão de relacionamentos, dos quais ligam os conceitos às hierarquias. Ela inclui sinais, sintomas, diagnósticos e procedimentos. Seu projeto único irá permitir a integração completa de todas as informações médicas em um registro médico eletrônico, contendo uma estrutura única de dados(Snomed, 2010). Em abril 2007, SNOMED CT foi adquirido perto IHTSDO (Organização Internacional do Desenvolvimento dos Padrões da Terminologia da Saúde).

A sistematização do modelo de dados SNOMED compreende uma combinação de alguns eixos para formular um diagnóstico. Por exemplo, um diagnóstico completo na SNOMED consiste em um código topográfico, um código morfológico, um código de organismo vivo e um código funcional. Quando houver a combinação desses quatro códigos para formação de um diagnóstico é estabelecido um novo código de diagnóstico. Por exemplo, a doença com o código D-13510 Pneumonia pneumocócica é equivalente à combinação de: T-28000 (código topográfico para pulmão); M-40000 (código morfológico para

inflamação) e L-25116 (código para Streptococcus pneumoniae do eixo de organismos vivos).

A Figura 7 apresenta um aplicativo que permite a manipulação do SNOMED. Nessa tela é possível visualizar os principais eixos da ontologia SNOMED, bem como as descrições detalhadas de cada termo, e uma lista dos itens da hierarquia, que estão mais próximos do selecionado.

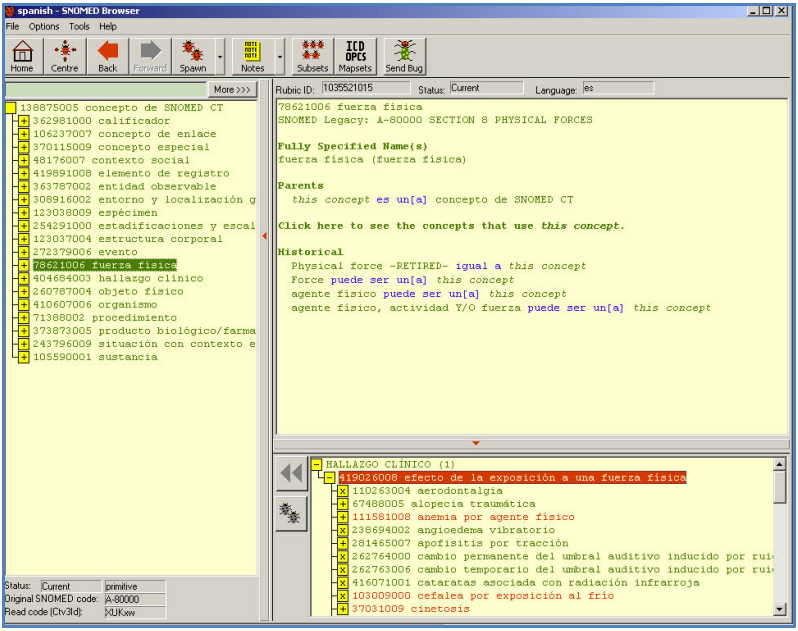


Figura 7: Exemplo da hierarquia do Snomed.
Fonte: (Snomed, 2010).

2.2.2 MESH

O vocabulário MeSH (*Medical Subject Headings*) é um vocabulário hierárquico desenvolvido pela *National Library of Medicine* (USA), que contém cerca de 36.000 conceitos médicos, abrangendo os mais diversos assuntos.O MeSH é um *thesaurus* que funciona como um dicionário. É composto por uma lista de palavras ordenadas e organizadas por tópicos ou contextos. Foi desenvolvido como um projeto da Biblioteca Nacional Americana. Todos os livros e artigos

publicados em medicina são catalogados e indexados de acordo com esse índice.

O MeSH também é a base de indexação do *Medline*, que é um sistema para pesquisa e para recuperação de literatura publicada na área médica. O MeSH tem uma vasta cobertura dos tópicos e apresenta um total de 16.148 verbetes e 73.641 sinônimos e variantes léxicas. Os tópicos presentes no MeSH são:

A: Termos Anatômicos
B: Organismos
C: Doenças
D: Medicamentos
E: Técnicas e Equipamentos Analíticos, Diagnósticos ou Terapêuticos
F: Psiquiatria e Psicologia
G: Ciências Biológicas
H: Ciências Físicas
I: Antropologia, Educação, Sociologia, e Fenômenos Sociais
J: Tecnologia, Indústria, Agricultura e Alimentos
K: Humanidades
L: Ciência da Informação e Comunicação
M: Grupos de Pessoas
N: Saúde
Z: Geografia e Dados Geográficos

Cada categoria é dividida em subcategorias. Dentro de cada subcategoria, descritores são ordenados hierarquicamente do mais geral para o mais específicos em até 11 níveis hierárquicos. Essa árvore pode ser representada como mostrado em um sistema de classificação oficial, mas especialmente como um descritor de guias e utilidades para as pessoas que estão procurando títulos para documentos ou pesquisando em literaturas. A árvore não é uma classificação exaustiva do assunto, mas contém esses termos que foram selecionados para inclusão nesse *thesaurus*. Sua estrutura representa frequentemente um compromisso entre o cenário e a necessidade de disciplinas particulares e usuários, na ausência de alguma simples universalidade aceita acomodada.

Por causa da ramificação da estrutura das hierarquias, essa lista é algumas vezes referenciada como “arvores”. Cada descritor MeSH aparece ao menos em um lugar na árvore e pode aparecer em outros lugares se for o caso. Caso queira, o usuário pode navegar entre as árvores para encontrar títulos de assuntos adicionais mais específicos que um determinado título, e títulos mais amplos. Por exemplo, abaixo de ANOMALIA, têm-se as específicas anomalias:

Anomalia C16.131
Anomalia, Medicamento induzido C16.131.42
Anomalias, Múltiplas C16.131.77
Síndrome de Alagille C16.131.77.65
Síndrome de Angelman C16.131.77.95

Nessas árvores, cada descritor é seguido por um número que indica a localização nessa árvore. Também pode ser mostrado por um ou mais números, em menores gêneros, e truncados até o terceiro nível, indicando outras localizações da árvore do mesmo termo. O número serve somente para localizar o descritor em cada árvore e para ordenar aquele em determinado nível da árvore. Os números têm um significado interno. Por exemplo, *D12.776.641* e *D12.644.641* ambos estão no grupo 641 da árvore, mas não implica nenhuma característica comum. "D (Medicamentos), D12 (proteínas, peptídeo e aminoácidos) *D12.776.641* (proteína do tecido nervoso) e *D12.644.641* (peptídeo cíclico)".

2.2.3 DeCS

O DeCS (Descritores em Ciências da Saúde) foi criado, em 1986, pela Bireme a partir do MeSH que, por sua vez, surgiu em 1963. O DeCS destaca a importância do vocabulário estruturado (Decs, 2010):

> Vocabulários estruturados são necessários para descrever, organizar e prover acesso à informação. O uso de um vocabulário estruturado permite ao pesquisador recuperar a informação com o termo exato utilizado para descrever o conteúdo daquele documento científico. Os vocabulários estruturados funcionam também como mapas que guiam os usuários até a informação. Com a expansão da Internet, e o número de potenciais pontos de acesso à informação crescendo exponencialmente, os vocabulários podem ser úteis provendo termos consistentes que permitam ao usuário selecionar a informação que necessita a partir de uma vasta quantidade de dados.

Pode-se dizer que o DeCS é uma versão traduzida do MeSH, utilizado pelas fontes de informação que compõem a Biblioteca Virtual em Saúde e incorpora outras quatro categorias hierárquicas, a saber: *Homeopatia, Saúde Pública*, *Ciência e Saúde* e *Vigilância Sanitária*.

Ele foi desenvolvido com o objetivo de permitir o uso de uma terminologia comum para pesquisa em três idiomas (português, inglês e espanhol) e proporciona um meio para recuperar a informação independente de idioma, pois um texto procurado em português pode retornar o mesmo descritor em inglês e espanhol e vice-versa. Possui um vocabulário dinâmico com 30.369 descritores, dividido em 20 hierarquias. A Figura 8 apresenta um exemplo de uma categoria sendo visualizada em uma árvore hiperbólica.

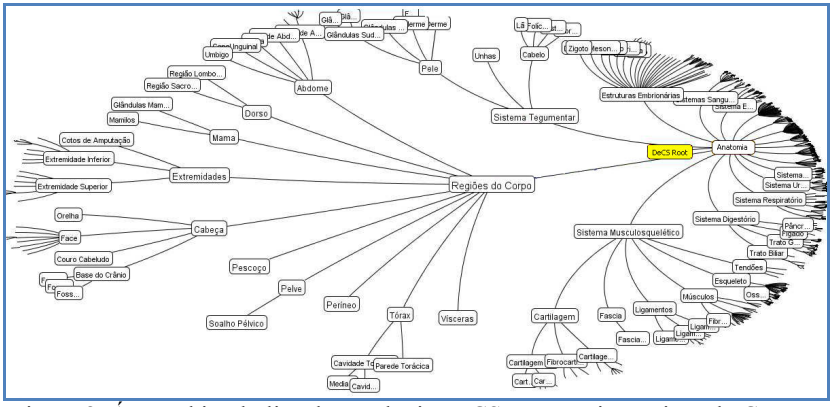


Figura 8: Árvore hiperbólica da ontologia DeCS - categoria Regiões do Corpo. Fonte: (Fileto, 2011).

2.2.4 UMLS

A fim de representar o conhecimento médico em um nível mais complexo e promover a integração dos diversos sistemas existentes, a Biblioteca Nacional Americana iniciou em 1989 o Projeto UMLS (*Unified Medical Language System* - Sistema de Linguagem Médica Unificada). Esse é um projeto que envolve diversas universidades, em um esforço para unificar e mapear todos os vocabulários existentes a fim de criar um *metathesaurus* (Bodenreider e Burgun, 2004).

A UMLS incorpora dentre outros, os seguintes vocabulários: CID, SNOMED, MeSH, GO (*Genome Annotations*), LOINC. Todos os termos são listados em um *thesaurus* e relacionados por meio de uma rede semântica. Existe um mapeamento entre os diferentes vocabulários que permite a identificação e a codificação de termos segundo diversos

sistemas. A UMLS é extremamente complexa e serve para diversos propósitos nas áreas de assistência, de ensino e de pesquisa.

Para serem intercambiáveis entre instituições diferentes e para possibilitarem pesquisas epidemiológicas, sugere-se que os registros em saúde sigam uma nomenclatura comum. Os elementos de dados, idealmente, devem ser compatíveis entre sistemas. Algumas iniciativas sugerem a utilização de conceitos baseados no UMLS, ou esses conceitos devem ser mapeáveis para o UMLS. A Figura 9 apresenta um exemplo do *metatheasurus* UMLS com as diversas ontologias que fazem parte do projeto.

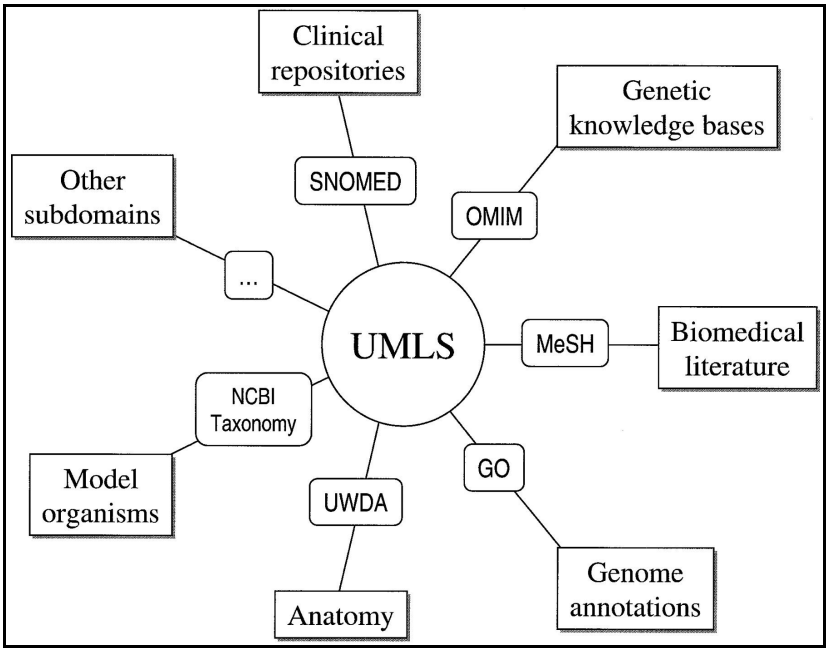


Figura 9: *Metasthesaurus* UMLS
Fonte: (Bodenreider e Burgun, 2004).

A UMLS possui três fontes de conhecimento: um *Metathesaurus*, composto por mais de 1.000.000 de conceitos distribuído em 17 diferentes idiomas, mapeados na forma de um *thesaurus*; uma rede semântica (*Semantic Network*), que possui uma estrutura similar a uma ontologia, onde 135 categorias principais são organizadas juntamente com 54 relacionamentos entre essas categorias; e uma caixa de

ferramentas especialistas (*SPECIALIST Lexicon & Tools*), que contém informações e recursos para serem utilizadas em PLN.

Os recursos da UMLS estão disponíveis em um serviço chamado de *UMLS Knowledge Source Server,* que está disponível no site do fabricante[2].

Esse servidor é um *Metathesaurus* (um banco de dados), que contém um vocabulário multilíngüe, contém informações sobre conceitos que dizem respeito às áreas da saúde, suas denominações e relacionamentos entre os termos. Ele pode ser utilizado para múltiplos fins, dentre os quais, na assistência ao paciente, faturamento de serviços de saúde, estatísticas em saúde pública, indexação e criação de catálogos de literatura biomédica, ou em pesquisas básicas sobre serviços de saúde. A UMLS está organizada por conceitos ou significados, que estão associados e agrupados a nomes alternativos e a diversas visões de um mesmo conceito, de forma a permitir identificar as relações úteis entre os diferentes conceitos.

[2]https://login.nlm.nih.gov/cas/login?service=http://umlsks.nlm.nih.gov/uPortal/Login.

3 RECUPERAÇÃO DA INFORMAÇÃO

A maioria dos sistemas de RI tem como objetivo recuperar a informação de forma a corresponder as palavras-chave do usuário em sua pesquisa (Singhal, 2001; Manning *et al.*, 2008). A possibilidade de um sistema recuperar os documentos relevantes à intenção do usuário é extremamente limitada, pois o maior problema é saber se a resposta está correta ou não (Van Rijsbergen, 1975). Normalmente um usuário refaz sua pesquisa à medida que recebe as primeiras informações ou então os documentos recuperados satisfazem em parte a necessidade do usuário (Spink *et al.*, 2001). Por esse motivo, o conjunto de documentos correspondentes deve estar o mais próximo possível das palavras utilizadas pelo usuário.

Atualmente existem várias abordagens de tecnologias de RI que foram propostas ao longo de vários anos, mas as principais são a clássica Booleana, o modelo vetorial, o modelo probabilístico e os modelos baseados em PLN. Na abordagem clássica, os documentos que são recuperados são baseados em *qeuries* que utilizam operadores Booleanos. A abordagem que utiliza regras de PLN tenta implementar a análise sintática e a semântica para reproduzir uma compreensão dos textos da linguagem humana (Greengrass, 2001).

Os primeiros sistemas de RI desenvolvidos foram os booleanos, que permitiam aos usuários especificarem as informações do que necessitavam utilizando a combinação dos operadores AND, OR e NOT (Manning *et al.*, 2008). A precisão nas respostas depende exclusivamente da ação do usuário, que muitas vezes é dificultada pela inexperiência na formulação da pesquisa. Consequentemente, uma resposta, nesse modelo, pode conter muito ou pouco resultado. Para melhorar os critérios de busca, os sistemas de RI utilizam modelos Booleanos híbridos, concatenados com outras técnicas, como por exemplo, (Savoy, 1996; Brin e Page, 1998; Jones *et al.*, 1998; Silverstein *et al.*, 1999; Kotsakis, 2002; Sauvagnat *et al.*, 2006).

A necessidade em melhorar os índices de respostas motivou a criação da segunda abordagem, que utiliza a semântica para recuperar a informação. Esse modelo leva em consideração a proximidade de termos, o uso de dicionários léxicos e semânticos e, as técnicas de PLN, para processar uma pesquisa. Normalmente esse modelo é mais utilizado em bases de dados estruturadas ou semiestruturadas (documentos XML). Em geral, é necessário especificar um caminho, como por exemplo, nome do banco de dados, ou uma tabela para que o

motor de busca possa identificar os documentos. O nível semântico tenta interpretar sentenças e não somente palavras individuais. Os métodos semânticos podem melhorar significativamente as técnicas de normalização, pois os termos podem ser mapeados em um padrão de sintaxe e armazenados em um formulário de índice (Greengrass, 2001). A seguir são apresentados os modelos mais utilizados para recuperar a informação na área de RI e que estão relacionadas com o escopo dessa pesquisa.

3.1 MODELO BOOLEANO

É o método mais tradicional de pesquisa. Possui baixa complexidade e é a forma mais utilizada pelos usuários para recuperar uma informação (Manning *et al.*, 2008). O modelo consiste em uma busca por todas as expressões em forma booleana, ou seja, os termos são indexados por meio da combinação dos operadores AND, OR e NOT. O método é eficaz na recuperação de documentos baseados em palavras simples. Pois o modelo vê cada documento como um conjunto de palavras. Nesse modelo é possível utilizar aspas para termos compostos e parênteses para definir prioridades dos operadores.

Para a recuperação da informação, esse modelo usa o conceito de índice invertido. O índice invertido é um método de fatorar as palavras do texto e guardar somente as referências das ocorrências nos documentos relevantes (Strohman, 2007). É criada uma lista ordenada com palavras-chave que contem para cada palavra, um apontador para cada um dos documentos onde a palavra aparece. Ele guarda também a posição da palavra no documento (Van Rijsbergen, 1975). Um exemplo de índice invertido é mostrado na Figura 10.

Os passos para recuperar uma informação em modelo booleano são:

1) Selecionar os documentos;
2) Simbolizar o texto: transformar cada documento em uma lista de símbolos.
3) Processamento lingüístico: normalizar o texto. Utilizar somente o radical das palavras, retirando os prefixos e sufixos (*stemming) (Strohman, 2007)*, retirar as *stop words*, converter o texto para caracteres em minúsculos.

4) Indexar os documentos com os termos em ocorrência para criar o índice invertido. Como resultado, um dicionário e uma lista de posições.

Sistema		(B, 3)	(C, 12)
Informática	(A, 10)	(B, 6)	(C, 2)
Médica	(A, 12)	(B, 6)	
Buscar	(A, 3)		

Figura 10: índice invertido.

Nesse exemplo, a palavra *Informática* aparece no documento *A* dez vezes, no Documento *B*, seis vezes e no documento *C*, duas vezes. Já a palavra *Buscar* aparece três vezes apenas no documento *A*.

- **Vantagens**: Utilizando o índice invertido, a busca torna-se bem mais eficiente. As consultas são simples e fáceis de entender. O modelo permite a execução de consultas estruturadas e relativamente fácil de desenvolver.
- **Desvantagens**: Um arquivo de índice invertido pode ter de 10 a 100% do tamanho do texto que deu origem ao índice (William e Ricardo, 1992). Isso acontece porque um índice invertido é definido por uma lista ordenada de palavras-chave, que contém para cada palavra, um apontador para cada um dos documentos em que a palavra ocorre e a posição da palavra nesse documento. Como resultado, à medida de um índice cresce em diversidade de termos, mais rápido fica o processo de indexação de novos termos, em conseqüência, mais espaço é ocupado no índice.

A maioria dos modelos de busca aqui pesquisados utiliza o modelo booleano (Savoy, 1996; Brin e Page, 1998; Jones *et al.*, 1998; Silverstein *et al.*, 1999; Kotsakis, 2002; Sauvagnat *et al.*, 2006). Como esse modelo possui muitas deficiências, como por exemplo, não possui *ranking,* a precisão nas respostas depende exclusivamente da ação do usuário, que muitas vezes é dificultada pela inexperiência na formulação da pesquisa e conseqüentemente, uma resposta pode conter muito ou pouco resultado. Mas para melhorar os critérios de busca, todos os sistemas de RI utilizam modelos Booleanos híbridos, concatenados com outras técnicas.

Por exemplo, Savoy (1996) avalia várias soluções existentes na época e apresenta uma tabela de precisão das buscas. O autor comparou os modelos de espaço vetorial, probabilísticos e o modelo booleano. Os

sistemas que retornaram os melhores resultados foram os que não utilizam modelo Booleano.

3.2 MODELO DE ESPAÇO VETORIAL

O modelo de espaço vetorial é representado por um vetor de termos, onde os documentos e as consultas possuem vetores associados (Singhal, 2001). O conjunto de termos utilizados é chamado de vocabulário. Os termos são ocorrências. O objetivo do modelo vetorial é avaliar o grau de similaridade entre um conjunto de palavras que compõe a busca e o documento. Para um termo ser similar a outro, o modelo vetorial associa pesos não binários aos termos de índices do conjunto de documentos. Esses pesos especificam o tamanho e a direção do vetor de representação, ao qual determinam a relevância de cada termo referenciado às consultas dos usuários.

Os documentos recuperados pelo modelo vetorial são ordenados decrescentemente ao grau de similaridade e o modelo leva em consideração que os documentos recuperados satisfazem parcialmente a busca. Desta forma, o conjunto de respostas é bem mais preciso do que os documentos recuperados pelo modelo booleano (Manning *et al.*, 2008). Se um termo pertence a um texto, o modelo define um valor diferente de zero no vetor correspondente ao termo. A maioria dos sistemas baseados em vetor opera no quadrante positivo do vetor espacial, ou seja, nenhum termo é atribuído como negativo. Para atribuir uma pontuação numérica para um documento em uma consulta, o modelo mede a similaridade entre o vetor de consulta e o vetor do documento. A similaridade entre os dois vetores faz parte do modelo (Faloutsos e Oard, 1995).

A principal vantagem desse modelo é a facilidade em computar a similaridade dos termos com eficiência e o fato de funcionar muito bem com sistemas genéricos. Uma desvantagem nesse modelo é que ele pode recuperar documentos que satisfizerem parcialmente uma consulta. O modelo faz um *ranking* dos documentos mais similares e ordena em ordem decrescente.

3.3 MODELOS PROBABILÍSTICOS

Os documentos são ordenados em uma coleção para diminuir a probabilidade de relevância em uma consulta. Nesse modelo, os termos indexados das consultas e dos documentos não são pré-definidos. A classificação é calculada por meio da geração dinâmica de pesos binários aos termos de uma pesquisa em relação aos documentos indexados. A classificação é gerada baseada no cálculo da probabilidade de um termo ser relevante a uma consulta. Esse modelo procura estimar a probabilidade de um usuário em encontrar um documento relevante e isso depende somente das representações do documento e das consultas. A ideia é que existe um conjunto ótimo de documentos que identifica a máxima relevância para o usuário. Os documentos que estão fora deste conjunto são considerados irrelevantes (Singhal, 2001).

A estimativa é a parte fundamental do modelo e, é onde a maioria dos modelos probabilísticos se difere uns dos outros. Muitos modelos probabilísticos têm sido propostos desde que o principal modelo foi desenvolvido por Maron e Kuhns (1996) e cada modelo é baseado em uma estimativa diferente de probabilidade técnica.

A principal vantagem desse modelo está em sua capacidade de construir uma função de classificação decrescente segundo a probabilidade de ser relevante a uma consulta (Jones *et al.*, 1998). A maior desvantagem desse modelo é que não há como saber quais documentos são relevantes a uma pesquisa. Desta forma, o modelo deve ser estimado inicialmente por interações com os usuários. Como esse modelo define os pesos de um termo por meio de atribuição binária, o modelo não processa a ocorrência de um termo em um documento. A maior desvantagem nesse modelo é a necessidade de segmentar uma coleção de documentos relevantes e irrelevantes sem considerar o número de ocorrências que os termos aparecem nos documentos (Maron e Kuhns, 1960).

3.4 MODELO DE INFERÊNCIA DE REDE

Nesse modelo, a recuperação de um documento é modelada como um processo de inferência em uma rede de inferência. O modelo de inferência de rede tem a capacidade de realizar uma classificação a partir de muitas fontes de evidências executados pela combinação dessas evidências. O modelo é basicamente uma rede Bayesiana usado para

modelos de documentos, conteúdos de documentos e para pesquisas (Turtle e Croft, 1990). Ela é consiste em dois métodos (Figura 11):

1. Documento de Rede, que é construído durante a indexação e é uma estrutura que não se altera durante o processo de busca e
2. Consulta de Rede, que é construído a partir de uma busca durante a recuperação da informação.

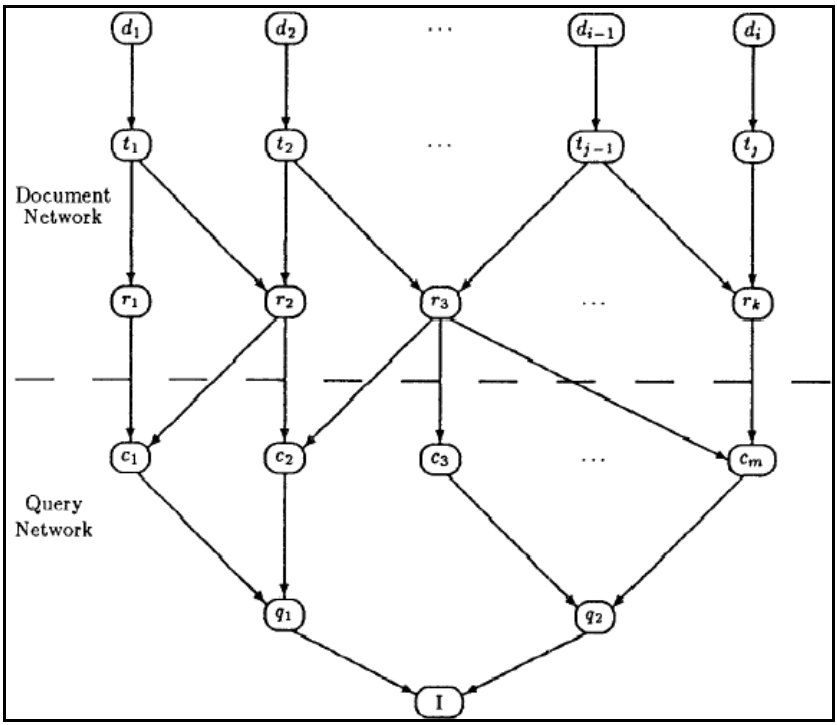


Figura 11: Modelo de inferência de rede.
Fonte: (Turtle e Croft, 1990).

O documento de rede representa uma coleção de documentos e contém um nó para cada documento e nodos para cada conceito de uma coleção. O nodo do documento representa as unidades de pesquisas na rede, isto é, contém os itens que se deseja visualizar como resultado da classificação. As setas (links) entre os nodos dos documentos e os conceitos dos documentos indicam que o conteúdo do documento é representado por um conceito. Cada link contém uma probabilidade condicional, ou um peso para indicar a força do relacionamento.

A consulta de rede representa a consulta enviada pelo usuário e consiste de um quadro de nodos que representam os conceitos necessários e os operadores, conectados em uma estrutura de árvore invertida. A consulta de rede é construída com um nó final (folha) que representa a necessidade de informação dos usuários. Esse modelo permite utilizar operadores booleanos, como AND, OR, NOT, SUM, MAX, dentre outros, para efetuar uma busca (Turtle e Croft, 1990).

A probabilidade de relevância documento é extraída no último nó e é usado para produzir a classificação. Nesse grafo significa que o primeiro nó tem influência direta sobre o nó subseqüente. Essa influência é apresentada por uma função de distribuição de probabilidade condicional que correlaciona os nós filhos com os estados dos nós pais. Desta forma, não é necessário utilizar grandes tabelas de probabilidade para todas as combinações de eventos possíveis.

Uma vantagem dessas redes é qualidade na representação dos relacionamentos probabilísticos. Nesse modelo é necessário calcular somente a independência entre as variáveis de um domínio. Essas independências são utilizadas para extrair o resultado de todas as variáveis que fazem parte da rede. O mecanismo de inferência é responsável pela busca das regras na base de conhecimento para serem avaliadas. Esse mecanismo direciona o processo de inferência de forma a efetuar uma boa interpretação e recuperar os objetos que estão representados na árvore de contexto. O mecanismo de inferência pode se tornar exponencial em alguns casos, mas é eficiente na maioria das situações onde é necessário recuperar uma informação.

3.5 RELEVANCE FEEDBACK

A operação chamada de *relevance feedbak* (realimentação por relevância) é um processo automático para alteração das requisições de pesquisas baseadas em avaliações de relevância solicitadas anteriormente pelos usuários na obtenção dos documentos (Orengo e Huyck, 2006). O usuário identifica, no conjunto de documentos inicialmente recuperados pelo motor de busca, um ou mais subconjuntos de documentos que julgar relevantes. O sistema extrai os termos comuns desse subconjunto de documentos e adiciona na expressão de pesquisa, de forma a refinar a busca (Daqing e Dan, 2011).

A ideia é que normalmente os usuários não são capazes de produzirem consultas perfeitas em uma primeira tentativa. O processo

melhora a especificação da consulta escolhendo termos importantes que foram anexados aos documentos obtidos anteriormente e que foram identificados como relevantes pelo usuário. Esse método consiste em solicitar aos usuários uma analise de uma amostra inicial de documentos recuperados para decidir se são relevantes. A consulta inicial é então modificada pelo sistema de RI e re-submetido ao motor de busca. Como resultado, uma nova lista de documentos recuperados é gerada.

Esse método, como é interativo, ele pode ser repetido várias vezes até que o usuário esteja satisfeito com os resultados (Orengo e Huyck, 2006). A principal vantagem desse modelo de pesquisa é que após a primeira resposta do sistema, o usuário interage com o mecanismo de busca não levando em conta o processo de formulação da pesquisa, identificando somente se os documentos são relevantes ou não. Outra vantagem desse método é a construção de um processo controlado que permite enfatizar alguns termos e diminuir a importância de outros.

3.5 LATENT SEMANTIC INDEXING (LSI)

O principal objetivo do modelo *Latent Semantic Indexing* – LSI é o mapeamento de cada vetor de termos indexados de um documento em um espaço dimensional reduzido (Rosario, 2000). A LSI tenta superar os problemas de correspondência léxica usando os índices conceituais ao invés de palavras individuais para a recuperação. Uma vez que existem diversas maneiras para exprimir um determinado conceito (sinônimos), os termos literais em uma consulta do usuário podem não corresponder às de um documento relevante.

Esse método presume que há uma estrutura semântica oculta (latente), subjacente aos dados. Nesse caso, a semântica é abandonada parcialmente pela aleatoriedade da escolha da palavra em referencia à recuperação, fato esse, que são escolhidas palavras individuais para serem recuperadas, e/ou indexadas.

Segundo Foronda (2005):

> O modelo matemático que se utiliza para criar a estrutura semântica é a decomposição Single Value Decomposition (SVD). O resultado da aplicação deste modelo, após realizadas operações matriciais, é uma matriz aproximada à matriz original. Esta matriz original é a matriz que representa uma relação, podendo ser esta relação

> termo- documento, termo-termo ou documento-termo. Matematicamente, este resultado pode ser interpretado como uma configuração espacial na qual o produto co-seno entre vetores representa a similaridade estimada entre dois documentos e, na área de RI, SVD é interpretada como uma técnica para gerar um conjunto de indexações não correlacionadas de variáveis ou fatores; cada relação (por exemplo, a relação termo-documento) é representada por seu vetor de valores.

Segundo Baeza-Yates (1999), a recuperação em um espaço reduzido pode ser melhor do que recuperar todo o espaço dos termos indexados. Por exemplo, em uma busca em que utilize a expressão "bases de dados bibliográficas em saúde", o sistema pode recuperar não somente as expressões que foram solicitadas, mas pode recuperar também os documentos que contenham os termos "pubmed" e "portal de teses e dissertações Fiocruz". Essas respostas são retornadas porque durante o processo de indexação, foi estabelecida uma relação entre os termos "pubmed" e "portal de teses e dissertações Fiocruz". E a cada novo documento que for indexado por esse modelo, o sistema adiciona novos termos de indexação e suas relações.

Uma das principais desvantagens dos sistemas LSI é eficiência, pois no modelo LSI as consultas devem ser comparadas com todos os documentos da coleção. Por exemplo, se a realimentação de relevância é realizada com o texto integral dos documentos pertinentes, o número de termos na consulta pode crescer muitas vezes o número de vetores LSI, levando a um aumento correspondente em tempo de pesquisa. O LSI funciona relativamente bem para documentos longos, devido ao pequeno número de vetores de contexto usados para descrever cada documento. No entanto, a implementação do LSI exige um investimento adicional de armazenamento e tempo computacional (Rosario, 2000).

3.7 EXPANSÃO DE PESQUISAS E ONTOLOGIAS

Para resolver o problema de uma pesquisa semântica, técnicas mais efetivas estão sendo desenvolvidas para analisar e recuperar semanticamente o conhecimento médico (Díaz-Galiano *et al.*, 2009), (Gschwandtner *et al.*, 2010), (Moskovitch e Shahar, 2009), (Bhogal *et al.*, 2007), (Gindl *et al.*, 2008). Nesses trabalhos aqui apresentados, as ferramentas utilizam técnicas de RI para analisar as pesquisas do usuário

e propor consultas mais amplas baseadas na idéia inicial. Um dos métodos mais comuns para recuperar uma informação é a utilização de respostas dos usuários como um modelo inicial e usá-lo para construir uma nova consulta a fim de se obter melhores resultados (*Relevance Feedback*).

Outra forma de melhorar as consultas é a utilização de diversas fontes de dados, como *thesaurus* ou ontologias, que são adicionadas à pesquisa do usuário (Grootjen e Van Der Weide, 2006). Técnicas que utilizam ontologias ou *thesaurus* são mais utilizadas em sistemas que recuperam informações semânticas, visto que são desenvolvidos procedimentos para o reconhecimento de expressões sinônimas e linguísticas, a fim de reconhecer textos semanticamente similares, mas sintaticamente distintos.

Um importante uso das ontologias é a expansão das pesquisas a partir de documentos escritos em texto livre com o objetivo de melhorar os resultados dos sistemas de RI (Munir *et al.*, 2006). Para expandir uma pesquisa, o usuário é guiado a reformular a pesquisa por meio da adição de termos mais significativos à pesquisa inicial. O sistema extrai os termos similares a partir de uma ontologia e gera uma nova pesquisa expandida. Os termos resultantes podem fornecer novas informações contextuais a partir da pesquisa inicial melhorando assim, a qualidade dos resultados (Bhogal *et al.*, 2007).

3.8 DETECÇÃO DE EXPRESSÕES NEGATIVAS

A negação, freqüentemente utilizada em textos médicos, é o processo de reverter o sentido de uma frase (Chapman *et al.*, 2002). Ela é muito importante para a comunicação entre pessoas. Expressões negativas são usadas por profissionais médicos para prover diagnósticos de pacientes, doenças, procedimentos, etc. Um resultado negativo em um diagnóstico médico é tão importante quanto um resultado positivo para auxiliar o profissional médico em sua decisão. Por esse motivo, uma frase negativa pode invalidar a pesquisa em sistemas de RI.

O processo de detecção de expressões negativas requer muito conhecimento sobre a linguagem para identificar corretamente as expressões ou termos de um documento (Gindl *et al.*, 2008). Em consultas a grandes bases de dados médicos em texto livre, a presença de expressões negativas pode conter muitos resultados falso-positivos. Isso porque o especialista médico é treinado para incluir essas

expressões em seus laudos. Uma forma de filtrar esses resultados é a criação de um mecanismo de detecção de expressões negativas em bases de dados médicos e assim, tornar mais ágil o processo de indexação e classificação da informação para o usuário.

Recuperar expressões negativas é tão importante quanto a recuperar qualquer outra informação do banco de dados. Portanto, o motor de busca do sistema deve decidir se uma expressão contida no banco de dados será excluída ou incluída na pesquisa (Gindl *et al.*, 2008). Fases ou expressões negativas são comumente utilizadas por médicos para prover um diagnóstico de paciente e procedimentos médicos. Elas ajudam o especialista a identificar sintomas e doenças que ocorrem a partir de documentos médicos (Mutalik *et al.*, 2001). Por exemplo, em um documento médico é comum encontrar diagnósticos que contenham textos com expressões do tipo: "o paciente não possui hipertensão arterial", ou "o paciente tem pressão alta".

Estas expressões podem criar um problema para os tradicionais motores de busca. Se o motor detecta a primeira expressão como verdade, todos os documentos que contenham esse tipo de expressão serão recuperados e o resultado irá conter informações falsas. Devido a isso, o processo de detecção de negação exige muito conhecimento sobre o idioma para identificar corretamente palavras negativas ou termos de uma expressão. A detecção de declarações negativas nos bancos de dados médicos pode reduzir o espaço de busca, assim, tornar o processo de busca mais ágil.

Nesse trabalho serão apresentadas técnicas para encontrar automaticamente expressões negativas contidas em textos médicos, criar um corpus anotado e extrair informações do banco de dados e armazenar no repositório semântico as expressões utilizadas.

3.9 ANOTAÇÃO SEMÂNTICA E REPOSITÓRIO SEMÂNTICO

Uma anotação semântica permite que computadores possam entender a semântica de um conjunto de dados. Esse processo é mais usado na área da *web* semântica, uma iniciativa promovida pelo *World Wide Web Consortium* (W3C), que visa o processo de adicionar conteúdo e/ou metadados em páginas da *web* para oferecer automação, integração e reutilização de dados entre diversas aplicações (Agosti *et al.*, 2007). Para cumprir estas exigências, uma anotação deve ser

baseada em um modelo de domínio formal (por exemplo, uma ontologia).

Para o desenvolvimento dessa tese, especialistas da área médica serão responsáveis por anotar semanticamente os termos mais usados na área que estão armazenados no banco de dados. A seguir, serão definidas entidades mapeadas que são conectadas as descrições semânticas da ontologia. O resultado desse processo de associação é gravado em um repositório de índice semântico. Esse repositório utiliza ponteiros para referenciar os termos da ontologia e as entidades nomeadas no texto.

O repositório semântico (exemplificado na Figura 12) constitui de uma base de dados contendo informações de diferentes fontes que mantém os índices múltiplos para diferentes tipos de conteúdo e permite executar diferentes tipos de pesquisas. São similares aos tradicionais sistemas de gerenciamento de banco de dados, mas permitem o armazenamento, consulta e administração de dados estruturados ou semi-estruturados. Os repositórios semânticos oferecem fácil integração de dados diversos e mais poder de análise desses dados (Kiryakov *et al.*, 2003a).

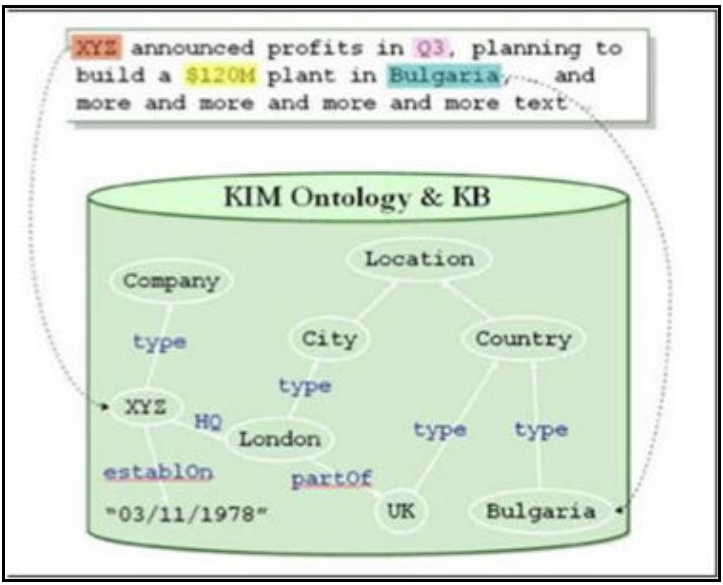


Figura 12: Anotação semântica utilizado por Kiryakov *et al*. (2003).
Fonte: (Kiryakov *et al.*, 2003b)

O repositório semântico desenvolvido para esse trabalho constitui de um *backbone* contendo informações das bases de dados de pacientes, laudos, exames, artigos científicos, toxicologia clinica, dicionários semânticos e ontologias médicas. O resultado da utilização de anotações semânticas juntamente com repositórios semânticos, consiste em basicamente um documento enriquecido com conhecimento compartilhado e maior liberdade de reutilizar em diversas pesquisas.

Uma anotação semântica fornece uma união entre os dados contidos em um documento e uma ontologia. Normalmente uma anotação semântica é uma referencia para um ou mais termos formalmente definidos em uma ontologia. Esse processo de anotar semanticamente um documento vem sendo estudado por várias iniciativas com o objetivo de tornar tal processo o mais automático possível, uma vez que realizar anotações semânticas de forma manual é muito trabalhoso (Kiryakov *et al.*, 2003a; Reeve e Han, 2005; Embley *et al.*, 2006). O processo de anotação semântica é definido pela criação de entidades nomeadas no texto, como por exemplo, nome, cargo, especialidade.

Como descrito anteriormente, o repositório semântico é responsável pelo armazenamento de informações médicas de diferentes tipos de bases de dados, que permite ao usuário executar tarefas e criar novas informações.

4 ESTADO DA ARTE

O trabalho aqui apresentado defende a utilização da técnica de expansão de pesquisa, juntamente com detecção de expressões negativas e anotação semântica de textos a fim de melhorar os índices de precisão nos procedimentos de busca em documentos médicos. Nesse sentido, serão apresentados os trabalhos relacionados ao tema dessa tese.

Foram utilizados os portais de pesquisa da *IEEEXplore*, *ACM Digital Library*, *ScienceDirect,* para recuperar os trabalhos correlatos a essa tese. A pesquisa dividiu-se em três categorias de busca por palavras chaves em cada uma dos portais:

1) Expansão de pesquisas. Nessa categoria foram pesquisadas nos portais as seguintes expressões: Query expansion, medical databases, ontology, indexing. Essa pesquisa retornou no ScienceDirect 66 artigos, no IEEEXplore, 132 artigos e no ACM Digital Library, 8 artigos. Nesse último portal de pesquisa, foram utilizados outros conjuntos de palavras chaves, mas a maioria já continha os mesmos trabalhos encontrados nos outros dois portais de pesquisa. Dessa forma, continuou-se com os mesmos critérios para os três portais de pesquisa. Desse conjunto de artigos encontrados (206), foram selecionados cinco mais importantes que tratavam somente sobre pesquisas em bases de dados biomédicos e que continham pesquisas referente à ontologias e indexação de dados usando ferramentas de buscas.
2) Detecção de expressões negativas. Aqui foram pesquisadas nos portais de buscas as expressões: negation detection, natural language processing, information retrieval e clinical data mining. Essa pesquisa retornou no ScienceDirect 51 artigos, no IEEEXplore, 48 artigos e no ACM Digital Library, 53 artigos. No portal de pesquisa IEEEXplore, foram utilizados outros conjuntos de palavras chaves, concatenadas com operadores OR e AND, mas a maioria já continha os mesmos trabalhos encontrados nos outros dois portais de pesquisa. Dessa forma, continuou-se com os mesmos critérios para os três portais de pesquisa. Desse conjunto de artigos encontrados (153), foram selecionados dez mais importantes que tratavam somente sobre técnicas de extração de termos em bases de dados médicas e que tinham relação direta com informações de saúde. Houve casos onde dois ou mais artigos

selecionados referenciavam a um mesmo estudo. Nesse caso somente um dos artigos foi selecionado. Ao final foram detalhados somente os três mais relevantes ao objetivo dessa pesquisa;

3) Anotação semântica de textos. Nessa última categoria de pesquisa foram executadas as seguintes expressões: Semantic Annotation, Medical Databases, Indexing, Medical Ontology, Knowledge, Semantic Search. Essa pesquisa retornou no ScienceDirect 141 artigos. Utilizando os mesmos critérios de buscas no IEEEXplore, foram encontrados somente 2 (dois) artigos que satisfaziam a busca. Nesse caso foram incluídos o operador OR à expressões para expandir o universo de pesquisa. Com a nova expressão (Semantic OR Annotation OR Database OR Medical OR Ontology OR Knowledge OR Semantic OR Search) foram encontrados 9(nove) artigos e no ACM Digital Library,foram encontrados somente 9 (nove) artigos. Nesse último portal de pesquisa, foram utilizados outros conjuntos de palavras chaves juntamente com o operador lógico OR para compor a nova pesquisa (semantic annotation, indexing medical data, OR medical ontology). Nesse cenário, foram encontrados 121 artigos. Do total de artigos encontrados (271), foram selecionados cinco mais importantes que objetivam anotar semanticamente os termos e criar as anotações semânticas (entidades nomeadas) em bases de dados médicas baseadas em ontologias.

Ao final da revisão sistemática, foram selecionados 11 (onze) artigos (apresentados em resumo no **Erro! Fonte de referência não encontrada.**), que satisfizeram todos os critérios de inclusão e exclusão descritos acima e, que serão de base referencial ao desenvolvimento dessa tese.

Para cada um dos artigos estudados será apresentado um pequeno resumo, os resultados positivos e negativos encontrados e uma justificativa sobre como ele é importante para o desenvolvimento desse trabalho. O **Erro! Fonte de referência não encontrada.** mostra um resumo dos artigos encontrados e quais os parâmetros que foram utilizados em cada um dos portais pesquisados.

Indexador	Parâmetros utilizados	Artigos retornados
Expansão de pesquisas		
ScienceDirect	*Query expansion, medical databases, ontology, indexing*.	66
IEEEXplore	*Query expansion, medical databases, ontology, indexing*.	132
ACM Digital Library	*Query expansion, medical databases, ontology, indexing*.	8
Detecção de expressões negativas		
ScienceDirect	*negation detection*, *natural language processing*, *information retrieval* e *clinical data mining*	51
IEEEXplore	*negation AND detection*, OR *natural AND language AND processing*, OR *information AND retrieval AND clinical data mining*.	48
ACM Digital Library	*negation detection*, *natural language processing*, *information retrieval* e *clinical data mining*.	53
Anotação semântica de textos		
ScienceDirect	*Semantic Annotation, Medical Databases, Indexing, Medical Ontology, Knowledge, Semantic Search*	141
IEEEXplore	*Semantic OR Annotation OR Database OR Medical OR Ontology OR Knowledge OR Semantic OR Search*.	9
ACM Digital Library	*semantic annotation, indexing medical data, OR medical ontology*	121
Total		**630**

Quadro 1 Parâmetros utilizados para pesquisa para as três categorias

4.1 EXPANSÃO DE BUSCAS

Díaz-Galiano *et al.*, (2009) apresentaram um método para expandir consultas usando ontologias médicas para melhorar os sistemas de RI. O objetivo é combinar dois subsistemas independentes para recuperar informações textuais e informações visuais usando a integração do conhecimento médico com expansão de termos. Esse método obteve melhores resultados usando as duas técnicas, quando comparado com cada uma das técnicas em separado. Nesse trabalho, os autores usam um corpus multimodal com informações médicas mantido pelo *Cross Language Evaluation Forum (CLEF – http://www.clef-campaign.org)*, para pré-processar a coleção de dados e extrair informações textuais associadas com cada imagem armazenada em seu banco de dados.

Para expandir as pesquisas textuais, os autores usaram o cabeçalho (de somente três categorias) e a lista de sinônimos do descritor MESH para adicionar a informação médica à pesquisa. Como resultado, ao se utilizar o conhecimento médico às pesquisas, os autores obtiveram melhores resultados quando não utilizado o conhecimento semântico. Quando testados somente as pesquisas textuais, o índice médio de *precision* foi de 0,1461 para as pesquisas sem conhecimento e de 0,2006 usando a expansão de pesquisas. Para esse exemplo, a melhoria chegou a 37,30% em pesquisas usando a ontologia MeSH. O modelo de expansão de pesquisa apresentado pelos autores pode ser visualizado na Figura 13.

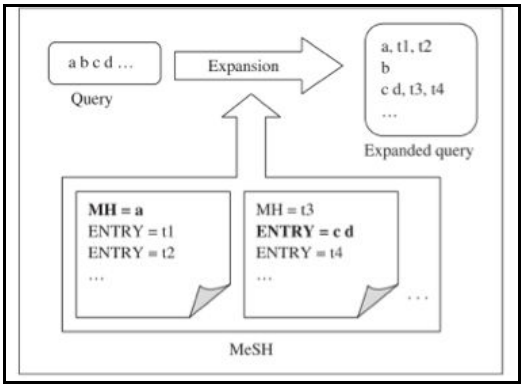


Figura 13: Exemplo de expansão de pesquisa
Fonte: (Díaz-Galiano *et al.*, 2009).

Lin *et al.,* (2005)(Lin *et al.*, 2005) apresentaram uma abordagem para recuperar artigos relevantes em um corpus biomédico. Os autores consideraram quatro tipos de operadores para expandir a pesquisa do usuário a partir da ontologia MeSH. Os operadores foram utilizados para definição dos pesos de uma consulta, priorizar e ranquear os resultados da pesquisa. Frases encontradas na base que contenham todos os termos têm pesos maiores que as que não contêm. A lista de sinônimos encontradas na ontologia foi utilizada para expandir a pesquisa do usuário.

Mas ao invés de definir uma única consulta, a abordagem utilizada pelos autores gera uma consulta para cada sinônimo encontrado na ontologia. Os resultados apresentados pelos autores mostram que as buscas utilizando expansão de pesquisas, reduziram a precisão em 25,43% quando comparado com uma consulta tradicional. Ao aplicar somente a priorização dos documentos, essa redução foi somente de 1,43%. Como resultado, os autores concluíram que a expansão da consulta reduz a desempenho do sistema, mas a priorização juntamente com a expansão da busca dos documentos melhorou os resultados da pesquisa que tinha somente a expansão da consulta.

Outros trabalhos estudam o problema da expansão da consulta, como forma de enriquecer os sistemas de recuperação de informação (Bhogal *et al.*, 2007), (Abdou e Savoy, 2008). Uma maneira de expandir a consulta é a utilização da técnica de inversão de freqüência de termos - *Term Frequency – Inverse Document Frequency* (TF-IDF). A fim de expandir a pesquisa, esse método inclui novos termos nas consultas usando palavras ou frases que possuem significados semelhantes ou relacionados com a pesquisa original (Abdou e Savoy, 2008).

Para avaliar a importância da expansão de pesquisa e indexação manual, Abdou e Savoy (2008) desenvolveram um novo modelo de expansão de pesquisas e avaliaram o desempenho de dez diferentes técnicas de RI, incluindo a probabilística, a linguagem e os modelos de espaço vetorial. Os autores realizam dois testes de RI (com e sem a ontologia MeSH) a partir de uma coleção de dados da *Medline*, a fim de medir o desempenho da recuperação. O método apresentado pelos autores que utiliza modelos probabilísticos é 170% a mais preciso quando comparado com o modelo clássico de espaço vetorial. Quando foi adicionado a ontologia MeSH, para indexar os artigos científicos, o modelo apresentado pelos autores, resultou em uma melhora de 8,4% na precisão do esquema de RI probabilístico.

Bhogal *et al.*(2007) apresentaram uma série de definições para a recuperação da informação focada no uso de contexto para a expansão da consulta. Eles discutiram o problema do uso de ontologias para uma série de tarefas de recuperação de informação e sua utilização na área de expansão da consulta. O objetivo foi fazer uma analise dos estudos de caso sobre ontologia de domínio independente e de domínio específico, com o objetivo de analisar as razões para o sucesso ou o fracasso da expansão de pesquisa baseado em ontologias.

Para os autores o sucesso em usar ontologias para expandir buscas depende de vários fatores: qualidade do modelo de conhecimento, familiaridade com o modelo de conhecimento e navegabilidade do modelo de conhecimento. De nada adianta utilizar um modelo que contenha centenas de megabytes de informação, mas que não se tenha uma ferramenta estável, coerente e resistente para interpretar essas informações.

O usuário também deve conhecer o modelo aplicado e esse deve ser de fácil utilização para que se possa navegar na ontologia, interagir com o sistema de busca e selecionar termos relevantes. Para maiores informações, consulte (Bhogal *et al.*, 2007).

4.2 RECUPERAÇÃO DE FRASES NEGATIVAS

Gindl *et al.*, (2008) desenvolveram um método para detectar e classificar a informações negativas encontradas nas diretrizes de prática clínica (CPGs) no nível sintático com informações gramaticais do idioma Inglês. Seus estudos mostram que elementos gramaticais são usados para decidir se uma frase é negativa ou não. A classificação da negação permite que o médico especialista possa decidir quais terapias ou opções de tratamento são mais indicadas aos pacientes. Seus métodos de detecção sintática melhoraram os valores de precisão e recall.

Um estudo anterior apresentado pelos autores identifica os conceitos da ontologia UMLS juntamente com um scanner léxico, a fim de reconhecer e classificar um grande conjunto de padrões negativos que ocorrem no texto (Mutalik *et al.*, 2001). Os autores desenvolveram um programa baseado na tecnologia existente para a execução de analisadores baseados em gramática livre de contexto. O objetivo é identificar negações adverbiais, preposicionais, adjetivos e verbos.

Os resultados de seu estudo mostram que o sistema apresentado teve um *recall* de 83,51% e um *precision* 67,49% na detecção de

expressões negativas em documentos médicos. Um dos maiores problemas para o algoritmo foi encontrar negações adverbiais, pois o índice de falsos positivos e falsos negativos ultrapassou os 90%. Isso se deve ao fato do algoritmo utilizado somente classificar uma expressão negativa analisando os três termos anteriores e posteriores em uma expressão. Para frases onde o termo negativo não está diretamente relacionado com os termos mais próximos, o algoritmo não consegue compreender na expressão.

O trabalho apresentado por Chapman (2001)(Chapman, 2001) objetivou a identificação automática de expressões negativas em resultados de exames em laudos médicos. Nesse trabalho foram avaliados o uso e a freqüência de frases negativas em relatórios médicos escritos em texto livre. O algoritmo é baseado na utilização de expressões regulares encontradas na ontologia UMLS para compor uma lista de 66 frases negativas. Os autores mostraram que 97% dos conceitos marcados como ausentes na UMLS, também foram considerados ausentes pelo sistema.

Os trabalhos apresentados acima empregam abordagens baseadas em engenharia do conhecimento, onde especialistas modelam regras e padrões que são normalmente utilizados para capturar características sintáticas e semânticas de um texto. A maior parte dos métodos utilizados provém da área de Processamento Natural de Linguagem (*Natural Language Processing* - NLP). Estes métodos são em sua maioria complexos e exigem mais trabalho. NLP são viáveis quando os textos aos quais se aplicam, forem regidos por regras gramaticais bem definidas (Rokach *et al.*, 2008).

Embora freqüentemente sejam encontrados erros ortográficos na base de laudos que foi testada, os textos armazenados são suficientemente corretos para serem processados por esses algoritmos.

4.3 ANOTAÇÃO SEMÂNTICA E PESQUISA SEMÂNTICA

Moskovitch e Shahar (2009) desenvolveram um motor de busca genérico que usa o conceito de pesquisa em base de dados e sensíveis ao contexto para melhorar a qualidade da pesquisa de texto dentro de documentos de diretrizes de prática clínica (Clinical Practice Guidelines- CPGs). Eles programaram consultas sensíveis ao contexto para permitir que os usuários especifiquem estrutura ontológica que foi especializada a partir de regras de conhecimentos para realizar a

pesquisa nos documentos em texto livre ou em um formato textual semi-estruturado.

A ferramenta desenvolvida consiste em uma camada onde os conceitos, sub-conceitos e super-conceitos do MeSH são indexados hierarquicamente. Por isso, um documento clínico pode ser classificado em muitos sub-conceitos dentro do mesmo conceito hierárquico. Essa múltipla indexação permite que usuários possam recuperar conhecimentos indexados por múltiplos conceitos. Por isso, são definidos pesos para cada resultado das *query* do usuário, os documentos são ranqueados e apresentados em uma lista ordenada. A pesquisa baseada no contexto foi definida em função de cada elemento hierárquico do modelo contextual. Cada elemento possui uma forma para que ele possa ser indexado, consultado e recuperado.

O resultado desse processo foi a criação de uma lista de entidades nomeadas (meta anotação), que contem os principais tipos de termos, descrições e tipos de pesquisas que podem ser efetuadas. Como resultado, as pesquisas baseadas em conceitos obtiveram melhores resultados em comparação com as pesquisas baseadas em textos completos ou em pesquisas sensíveis ao contexto. Os autores definiram que o melhor índice de *recall* foi o de 50% e para isso avaliaram diversos métodos. Os métodos de pesquisas baseadas em conceitos tiveram um índice de precisão de 0,50 e o melhor método foi o que utilizava os conceitos indexados em somente três níveis da ontologia.

Uma limitação do estudo é que as bibliotecas aplicadas precisam ser indexadas manualmente e há a necessidade de classificação manual para cada novo documento adicionado à biblioteca digital.

O objetivo dessa tese é usar a idéia aqui do algoritmo expandir a busca em várias sub-categorias da ontologia. Ainda, será desenvolvido um mecanismo de indexação automática para que quando da entrada de novos documentos à base, esses sejam indexados rapidamente.

Mykowiecka *et al.*, (2009) desenvolveram regras para a extração da informação em bases de dados médicas. Nesse trabalho foi criada uma ontologia especial que traduz os conceitos em dois modelos: um para representar as estruturas hierárquicas e outro, gramáticas dedicadas para processar documentos e preencher os *templates* fornecidos pelo modelo proposto. Os autores desenvolveram técnicas linguísticas para extrair a informação de tecidos mamários e diagnósticos patológicos em laudos de mamografias. Nesse trabalho foram criadas regras para extração de termos gramaticais como palavras ambíguas, expressões negativas, *tokenização* de textos e expressões anáforas.

Foram analisados 705 laudos de mamografia e o sistema extraiu os termos mais utilizados que foram definidos para as pesquisas. Como resultado, a metodologia conseguiu alcançar uma precisão média de 94,25% no reconhecimento de expressões nos textos médicos. Uma limitação conhecida nos sistemas baseados em regras é a necessidade de prever todas as maneiras possíveis de expressar a informação a ser exigida. Se a gramática não abrange todas as possibilidades, a precisão cai e isso reflete o fato de que algumas expressões usadas por médicos que não foram previstos no sistema apresentado.

Munir *et al.*, (2006) desenvolveu um modelo para recuperação de informação semântica a partir de bases de dados heterogêneas. Nesse trabalho são utilizadas técnicas de semânticas de reformulação de pesquisas em bases de dados biomédicas, baseadas em ontologias e descrições de fontes de dados heterogêneas. Os autores apresentam uma técnica para fusão de ontologias que foram construídas a partir de informações de ontologias distribuídas e que podem ser exploradas para expandir as consultas a fim de atender as necessidades dos usuários.

Esta abordagem é baseada na disponibilidade e na geração de ontologias para cada fonte de dados e também no uso de uma ontologia global que define a visão integrada e virtual da distribuição de fontes de dados heterogêneas. A ontologia resultante da fusão fornece uma representação unificada de todas as ontologias subjacentes, utilizada na geração das consultas e reformulações de pesquisas, que podem ser aplicadas na extração do conhecimento.

A metodologia aplicada por Munir *et al.*, (2006) não pode ser aplicada nesse trabalho, pois não prevê a construção de ontologias. Entretanto, o artigo apresenta como uma base de conhecimento pode ser utilizada para recuperar informações semânticas de bases de dados biomédicos e prover uma visão global da ontologia. A ideia será utilizada nesse trabalho com a finalidade de criar um repositório semântico centralizado com informações de pacientes e artigos científicos.

Já Gschwandtner *et al.*, (2010) apresentam um sistema de anotação semântica que mapeia conceitos de uma ontologia médica (UMLS) e gera textos médicos em formato livre. Foi customizado um sistema de anotação de páginas web para que o novo aplicativo pudesse compreender o domínio médico. A aplicação gera um mapa de conceitos médicos (*metadados*) a partir da terminologia e esses conceitos são anotados semanticamente nos documentos da base de dados. Os profissionais especialistas podem visualizar e corrigir todos os tipos de informações anotadas no documento. Esse trabalho mostra que o

mapeamento dos conceitos médicos da ontologia pode fornecer informações semanticamente precisas para processamento de textos e ajuda a eliminar a ambigüidade dos diferentes significados.

Lourenço *et al.*, (2010) identifica termos relevantes em documentos eletrônicos a partir do processo de reconhecimento de entidades nomeadas. O objetivo é anotar as ocorrências de classes biológicas a partir de resumos ou textos completos dentro da biblioteca *PubMed*. Esse modelo também apresenta um índice semântico dos documentos e termos encontrados. A técnica é usada para extrair informação a partir de bibliotecas médicas, pré-processar os documentos e aplicar um dicionário léxico para realizar o reconhecimento de entidades nomeadas. O estudo apresentado pelos autores permitiu reduzir significativamente o número de documentos irrelevantes sem a perda dos documentos relevantes.

4.4 RESUMO ESQUEMÁTICO

A

Tabela 1 apresenta um resumo das técnicas de conhecimento que foram utilizadas para a recuperação de documentos em domínio médico. Pela bibliografia analisada, pode-se concluir que as técnicas utilizadas somente abrangem uma e, às vezes duas das áreas de extração de informação no sentido de ampliar as consultas e melhorar os índices de precisão das pesquisas do usuário. A maioria dos trabalhos apresentados é embasada em ontologias para indexar o conhecimento médico e auxiliar os sistemas de RI. Todas as técnicas apresentadas ampliaram, mesmo que minimamente, os resultados de buscas quando comparadas com as técnicas tradicionais da RI.

Tabela 1: Resumo dos artigos pesquisados.

Artigo	**Expansão de pesquisas**	**Anotação semântica**	**Extração de Expressões Negativas**	**Ontologia Utilizada**	**Objetivo Principal**
(Díaz-Galiano *et al.*, 2009)	X			MeSH	Expandir consultas usando ontologias médicas.

(Lin *et al.*, 2005)	X			MeSH	Expandir consultas para classificar documentos médicos
(Abdou e Savoy, 2008)	X			MeSH	Avaliação de dez diferentes técnicas de RI para medir a precisão das respostas.
(Bhogal *et al.*, 2007)	X			-	Estudo de caso sobre o uso de ontologias de domínio independente e específico, para avaliar as metodologias de expansão de pesquisas baseadas em ontologia.
(Moskovitch e Shahar, 2009)		X		MeSH	Motor de busca para pesquisas em dados sensíveis ao contexto.
(Mykowiecka *et al.*, 2009)		X	X	-	Criação de regras para extração do conhecimento em bases de dados médicas.
(Munir *et al.*, 2006)	X	X		-	RI semântica a partir de bases de dados heterogêneas.
(Gschwandtner *et al.*, 2010)		X		UMLS	Anotação semântica para mapear conceitos de ontologia médica.
(Lourenço *et al.*, 2010)		X		-	Identificar termos relevantes para criar entidades nomeadas e aumentar a precisão das buscas em documentos da

					PubMEd.
(Gindl *et al.*, 2008)			X	MeSH	Detectar e classificar a informações negativas em documentos médicos.
(Chapman, 2001)			X	UMLS	Identificação automática de expressões negativas em resultados de exames em laudos médicos.
Modelo Proposto	**X**	**X**	**X**	**DeCs**	**Recuperação e comunicação do conhecimento médico.**

Conforme apresentado no capitulo 1, seção 1,7, o primeiro resultado esperado desse trabalho é a criação de uma base de conhecimento do domínio médio que utiliza técnicas de expansão de pesquisa utilizando-se a ontologia DeCs para esse fim. Para desenvolver o modelo de expansão de consultas, será utilizada uma adaptação do modelo proposto por Díaz-Galiano *et al.*, (2009)(Díaz-Galiano *et al.*, 2009). Ao invés de utilizar o MeSH como referência para encontrar os sinônimos, esse trabalho recuperará as informações de expansão das pesquisas a partir da ontologia DeCS. Ainda, serão utilizadas todas as classes do DeCS para anotar semanticamente os textos médicos contidos na base de dados. A forma de como serão anotados os textos diferencia do modelo proposto por Díaz-Galiano *et al.*, (2009), pois nesse trabalho de pesquisa as informações serão não intrusivas, ou seja, as anotações não farão parte do laudo, mas sim da base de conhecimento que terá um laudo referenciado nela. Ainda, serão somente utilizadas informações textuais para a criação da base de conhecimento.

A abordagem utilizada por Lin *et al.*, (2005)(Lin *et al.*, 2005) será utilizada em partes para o desenvolvimento dessa tese. Foi constado que uma pesquisa que utiliza termos obrigatórios e frases completas reduz

muito a precisão da pesquisa. Ainda, os autores dividiram as consultas conforme o número de sinônimos encontrados na base do MeSH. Nesse trabalho será utilizada somente uma única consulta para todos os sinônimos que serão encontrados na ontologia e não serão consideradas como obrigatórias frases completas. Entretanto, as frases que forem encontradas na ontologia terão pesos maiores que os termos em separado.

Nos trabalhos apresentados nessa seção, não são abordados a expansão de pesquisa que leva em consideração a hierarquia do DeCS ou de qualquer outra ontologia de domínio. A principal diferença desse trabalho para os anteriores aqui apresentados será a possibilidade de utilizar essa técnica juntamente com os temos sinônimos e itens relacionados para aumentar a quantidade de termos em uma consulta, mas sem diminuir a precisão dos resultados.

O terceiro objetivo desse trabalho prevê o desenvolvimento de uma técnica de detecção de expressões negativas em textos médicos. Dessa forma, esse trabalho prevê a adaptação do algoritmo apresentado por Chapman, (2001)(Chapman, 2001) e do algoritmo de Gindl *et al.*, (2008)(Gindl *et al.*, 2008) para identificar automaticamente expressões negativas contidas em textos médicos na língua portuguesa. Um fato a ser considerado nessa pesquisa, é que nenhum dos trabalhos de detecção de frases negadas, prevê a utilização de expressões hipotéticas nos textos a serem processados. E essa tese de doutorado, além de conhecer os textos médicos que utilizam frases negativas, será capaz de tratar termos que apresentam expressões hipotéticas e também detectar frases que tenham duplo sentido.

5 MODELO PROPOSTO

A natureza desse trabalho é uma pesquisa aplicada dirigida à utilização de modelos de recuperação de informação semântica em bases de dados médicos e a representação do conhecimento gerada a partir de um repositório semântico.

Embora a maioria destas obras aplicarem diferentes técnicas para recuperação de informações de documentos médicos, todas elas têm uma função muito especial: melhorar o processo de busca de dados médicos. É compreensível que não é fácil melhorar o processo atual, mas baseado na união destas três técnicas previamente apresentadas (anotação semântica, expansão de pesquisas e detecção de expressões negativas), essa tese pretende refinar a abordagem atual e apresentar novas técnicas mais precisas para a recuperação de informação médica.

A principal motivação para o desenvolvimento desse trabalho é a criação de um modelo de recuperação e comunicação do conhecimento médico gerado a partir do Sistema Catarinense de Telemedicina e Telessaúde - STT. Para isso, faz-se necessário a apresentação de alguns requisitos não funcionais para o desenvolvimento do modelo de busca.

O primeiro requisito é garantir para o usuário uma ferramenta de busca transparente, ou seja, a ferramenta não deve mostrar o funcionamento da ontologia e nem qualquer outra forma de descrição do conhecimento.

O segundo requisito é garantir ao usuário que as informações solicitadas sejam mais precisas que as atuais ferramentas de buscas. Além disso, o modelo deve ser totalmente implementado em plataforma web.

Com base nesses requisitos, foram estudadas as principais ferramentas de buscas em sistemas de RI que melhor atendem as necessidades de trabalho e que pudesse permitir sua extensão e ou adaptação. Foi escolhido o *Lucene* por apresentar maior facilidade de utilização, por ser uma biblioteca de código aberto e possuir implementações em diversas linguagens de programação.

5.1 LUCENE

O framework utilizado para recuperar a informação é o Lucene (Hatcher e Gospodnetic, 2004) e será utilizado como base tecnológica da pesquisa. O Lucene é uma biblioteca ou uma API (do inglês,

Application Programming Interface) que permite indexar e recuperar documentos armazenados em arquivos, páginas web, ou em bases de dados relacional. O Lucene está sob o domínio da Apache Fundation e originalmente foi desenvolvido na linguagem Java, mas existem diversas implementações em diferentes linguagens de programação. O Lucene permite indexar e pesquisar qualquer dado que possa ser convertido para formato texto (Hatcher e Gospodnetic, 2004).

O core da engine de busca do Lucene é composta por dois componentes principais: Indexação e Pesquisa. Para pesquisar grandes volumes de texto, deve-se primeiramente indexar o texto e converter para um formato que se permita procurar a informação. Esse processo de conversão é chamado de indexação e o resultado é chamado de índice. Um índice é uma estrutura de dados para acesso randômico que pode ser simplesmente um arquivo que contém a localização dos arquivos indexados. A Figura 14 apresenta a arquitetura básica do Apache Lucene.

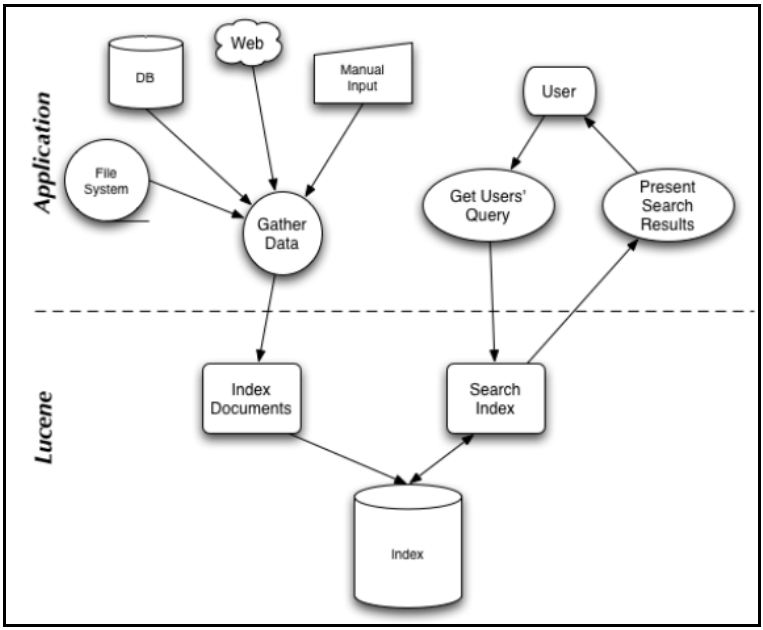


Figura 14: Integração entre a aplicação e o *Lucene*
Fonte: (Hatcher e Gospodnetic, 2004)

A pesquisa é o processo de procurar palavras em um índice. A qualidade da pesquisa é descrita utilizando métricas de *precision* (precisão) e *recall* (cobertura) onde: Recall define a quantidade de documentos relevantes recuperados entre os documentos relevantes existentes na base de dados. Para calcular o *recall* é necessário conhecer o total de documentos relevantes recuperados e o total de documentos relevantes que estão armazenados no Banco de dados.

Por sua vez, *precision* define a quantidade de documentos relevantes aos usuários dentre os documentos que foram retornados como resposta das buscas. Para calcular a *precision,* é necessário que o sistema saiba quais os documentos são relevantes na consulta e qual é a quantidade total de documentos que foram recuperados do banco de dados (Baeza-Yates e Ribeiro-Neto, 1999).

No entanto, um sistema de busca possui uma série de outros fatores para recuperar uma informação. Um sistema de busca deve retornar a informação rapidamente a partir de uma grande quantidade de dados. Deve ter suporte a pesquisas simples e múltiplas, frases longas e curtas, ranquear e classificar o resultado, para então apresentar ao usuário a lista de respostas que satisfazem a pesquisa.

O objetivo principal do *Lucene* é facilitar a recuperação da informação. Mas para se obter uma recuperação precisa é necessário que o texto seja analisado durante indexação. O *Lucene* analisa o texto para extrair termos separados em blocos. O processo de analise é a conversão do texto em termos. Esses termos são usados para determinar quais documentos correspondem a uma consulta durante as buscas. O método chamado de "analisador" realiza uma série de operações para facilitar o processo de indexação de um termo. Ele remove caracteres acentuados, converte as palavras para se ter somente caracteres minúsculos (processo chamado de normalização), remove as palavras comuns, como artigos e pronomes (*stop words*), extrai somente o radical das palavras (*stemming,* ou *tokenization*) ou muda as palavras para sua forma mais básica (lexema) (William e Ricardo, 1992; Manning *et al.*, 2008).

Por se tratar de um processo de análise de texto, é muito importante que o "analisador" dessa biblioteca tenha suporte a diferentes línguas, pois cada língua tem suas próprias características que são únicas. Outro fator a ser considerado é o domínio do texto a ser analisado. Na área médica utiliza-se uma terminologia bastante específica, siglas e abreviaturas que precisam de atenção especial. E um analisador simples não será suficiente para recuperar as informações de forma precisa (Alag, 2008).

Além do "analisador", o *Lucene* implementa as tradicionais consultas booleanas, utilizando os operadores "AND", "OR" e "NOT", pesquisas por expressões regulares que utilizam caracteres com funções especiais para formar uma busca (medic*, ou te?t) e pesquisas avançadas utilizando lógica *Fuzzy*. As pesquisas avançadas utilizam métodos de cálculo de distância (algoritmo de distância de Levenshtein (Gilleland, 2009)) para avaliar a proximidade e similaridade entre os termos de uma consulta. Esse processo permite classificar a informação em forma de *ranking,* para então apresentar ao usuário as respostas que mais satisfazem sua busca em ordem decrescente de relevância.

Entretanto, essa *engine* de pesquisa utiliza técnicas de busca tradicionais para indexar e analisar um documento. A utilização de modelos lingüísticos para indexar um conjunto de termos pode ampliar o *precision* e o *recall* da resposta, pois permite avaliar a distinção de termos similares não idênticos e estabelecer uma relação entre diferentes termos, fazendo com que termos com pequenas diferenças sejam identificados como um mesmo conjunto de termos. Os modelos lingüísticos estudam a morfologia, a sintaxe e a semântica para indexar um termo. Isso significa que no nível morfológico, a indexação é criada com o uso de *stemming*. No nível sintático, o objetivo é agrupar termos semanticamente equivalentes, mas sintaticamente diferentes (por exemplo, Pressão sanguínea alta, e Hipertensão). O nível semântico prevê a união de termos sinônimos (como "EGC", e "Eletrocardiograma"). Em um domínio específico, os termos sinônimos podem ser recuperados a partir de dicionários e *thesaurus* com o objetivo de melhorar o processo de busca.

Em um sistema com diferentes bases de informações, esse modelo não será suficiente para obter os melhores resultados. É preciso considerar a freqüência de múltiplos termos para poder representar de forma mais eficiente um sistema de busca. E uma forma de tratar múltiplos termos é a utilização de termos lexicais para analisar um documento. Ou seja, é necessário considerar palavras-compostas, nomes próprios, termos que se apresentam juntos, termos que estão separados por um determinado número de palavras e termos técnicos. Para se obter um melhor índice de *precision* e/ou *recall* é necessário recuperar a informação baseado também em conteúdo semântico (Moskovitch e Shahar, 2009).

5.2 ORGANIZAÇÃO CONCEITUAL: CAMADAS

Aqui estão organizadas as camadas do sistema desenvolvido. O sistema foi desenvolvido três camadas para facilitar a utilização e tornar o sistema mais flexível, permitindo que as partes possam ser alteradas de forma independente.

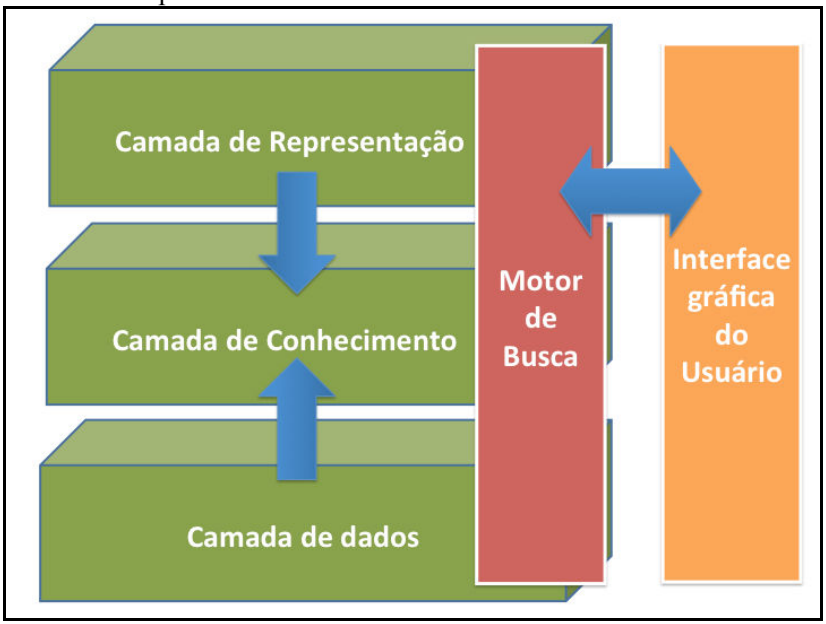


Figura 15: Arquitetura em camadas do modelo proposto.

A Figura 15 representa a arquitetura do modelo que foi dividida camadas, cada uma responsável por um tipo de informação. Esse modelo foi divido dessa forma para manter a compatibilidade com o STT.

A Camada de dados é o repositório de conteúdo do sistema. Nessa camada são armazenados os laudos da base do STT, o dicionário de sinônimos da língua portuguesa, o dicionário de expressões negativas, as expressões hipotéticas, e os termos utilizados pela linguagem cotidiana dos profissionais médicos. Nesse último caso, a inclusão de novos termos ao repositório deve ser validada por usuários especialistas.

A segunda camada do sistema é a de conhecimento. Essa camada corresponde à camada de ontologia e a base de conhecimento

propriamente dita. Aqui estão armazenadas as anotações semânticas (entidades nomeadas do DeCS e as expressões negativas encontradas nos laudos), a ontologia DeCS e suas relações com as instâncias dos conceitos que estão na camada de dados. A Figura 16 apresenta um exemplo das sentenças anotadas pelo modelo proposto. Por exemplo, na sentença 732 (no canto esquerdo abaixo da figura), é destacado o descritor "Hidrocefalia" da ontologia DeCs com um link para acesso à ontologia e também uma caixa de sugestão para indicar que a sentença 732 é uma expressão negada.

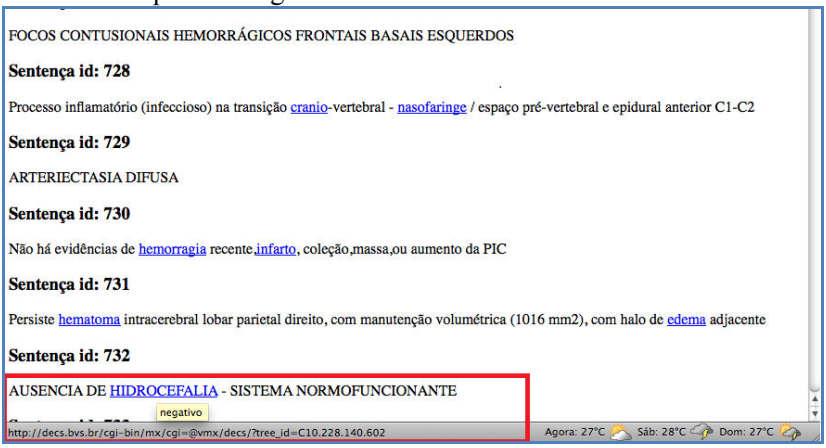


Figura 16: Exemplo de uma anotação apresentada pelo modelo proposto.

Essa ontologia deve ser conhecida pelos usuários especialistas do sistema para um melhor gerenciamento do sistema. Por exemplo, quando um usuário do sistema estiver procurando um determinado termo que não aparece na ontologia, mas que tem uma relação com algum termo da ontologia, o especialista poderá fazer a *linkagem* (conexão) desse termo com um termo relacionado na ontologia. Sua finalidade é servir de base para a anotação semântica dos conteúdos armazenados no repositório de dados.

Por fim, na camada de representação é responsável por expandir a descrição fornecida pela ontologia para introduzir os elementos do conhecimento dos usuários do sistema que não puderam ser representados pela ontologia. Por exemplo, o termo "calcificação em pipoca", que não é conhecido na ontologia, é descrito pela comunidade médica como uma linguagem cotidiana para o termo "neurocisticercose" constante na ontologia.

5.3 COMPONENTES DO SISTEMA

Essa seção apresenta os três principais componentes do modelo aqui proposto, o indexador, o mecanismo de busca e a interface gráfica do usuário. O desenvolvimento da interface gráfica não é estudado nesse trabalho, mas se faz necessário para melhor integração entre o usuário do sistema e o sistema propriamente dito. A interface gráfica foi desenvolvida no mesmo padrão do STT. Detalhes sobre o desenvolvimento do STT podem ser visualizados em (Wangenheim *et al.*, 2010).

5.3.1 Indexador

O indexador é o componente responsável por gerar os índices das informações contidas nas camadas do sistema. Após a definição dos campos que estão contidos na base de conhecimento o indexador percorre cada um dos documentos extraindo informações tais como as ocorrências de um termo, a sua localização, a freqüência com que o termo aparece no texto, dentre outros que permitam efetuar uma busca rápida sem que seja necessário examinar seqüencialmente toda a base de documentos. Como resultado desse processo, o índice é criado.

O primeiro processo para a geração do índice é chamado de analisador, que no caso da proposta dessa tese, foi utilizado como base o analisador do *Lunene* e adaptado às necessidades desse trabalho. Esse analisador é responsável por eliminar as *stop words,* efetuar a *tokenização,* definir o *stemming* para o termo, quando necessário, identificar as entidades nomeadas da ontologia e associar os termos da ontologia aos termos do banco de dados.

Após o processamento dessa informação, o índice é criado e começa a inserção dos documentos no índice. Esse processo é feito adicionando documento por documento que são lidos da camada de dados e cada campo que foi definido no processo anterior, é criado e populado. Maiores detalhes sobre a criação do indexador pode ser visto na seção 5.6 Indexação da base de conhecimento.

5.3.2. Motor de busca

O motor de busca utiliza as três camadas do sistema para efetuar uma pesquisa. As informações sobre a camada de representação são utilizadas para expandir semanticamente as buscas, incluindo os termos relacionados e os sinônimos da ontologia diretamente às palavras-chave inseridas pelo usuário e para definição da necessidade em buscar expressões negativas pelo motor de busca.

As informações definições da ontologia e da base de conhecimento são utilizadas para representar o conhecimento compartilhado pelos usuários, já o conteúdo utilizado pela linguagem cotidiana somente poderá fazer parte da busca após a validação do especialista. Quando da pesquisa por um termo que não é encontrado na ontologia e nem na base de conhecimento, esse termo será armazenado em um repositório temporário e quando validado pelo especialista em um segundo momento, poderá fazer parte da base de conhecimento e das buscas futuras.

5.3.3 Interface gráfica

Conforme descrito anteriormente, o desenvolvimento da interface gráfica do usuário não será abordado nesse trabalho. Nesse caso, foi desenvolvido um conjunto mínimo de funcionalidades somente para validação do modelo aqui desenvolvido. Essa interface foi desenvolvida seguindo os requisitos do STT e contêm somente os campos necessários para pesquisas por palavras-chave, uma área para a listagem dos resultados, um *checkbox* para pesquisas por expressões negativas, um botão para efetuar a pesquisa e um botão para adicionar termos desconhecidos à base de conhecimento (temporária).

Ao clicar no botão “adicionar à BC”, os dados são armazenados em uma base temporária. Quando o especialista for analisar essa base para validar os termos, o usuário é direcionado para outra interface que contém a visualização hierárquica da ontologia para ser relacionada com o termo inserido pelo usuário, conforme demonstrado na Figura 17.

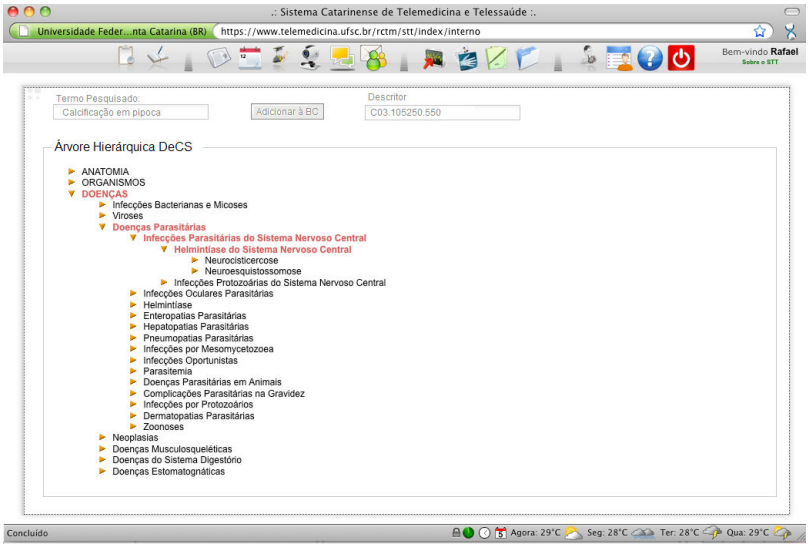


Figura 17: Tela de relacionamento para termo não encontrado na BC.

5.4 FUNCIONAMENTO DO SISTEMA

Para o desenvolvimento desse trabalho são utilizadas técnicas para recuperar informações textuais relacionadas à Ontologia DeCS e dicionários lingüísticos com o objetivo de ampliar o universo de pesquisa do usuário. Esse trabalho difere de trabalhos anteriores porque o objetivo é oferecer ao usuário uma série de documentos médicos muito mais amplo e eficaz. O objetivo é apresentar os documentos mais relevantes para que o usuário não precise despender muito tempo para encontrar a informação ou o usuário não precise procurar em bases de dados diferentes, a fim de encontrar a informação necessária. A abordagem apresentada aqui é baseada na técnica de expansão de pesquisa usando ontologias que define o uso de termos em árvore hierárquica e pelo uso de sinônimos. A fim de melhorar a pesquisa, anotações semânticas são utilizadas em textos médicos. Esta técnica inclui o uso de entidades nomeadas e detecção de frases negativas para aumentar o universo da pesquisa e reduzir o número de respostas menos relevantes. Para isso, foi criado um repositório de conhecimento que utilizara os conceitos da ontologia, a fim de extrair as informações dos documentos médicos e tornar o texto pré-processado e enviar os resultados ao usuário.

Quando um usuário especifica uma consulta, o motor de busca efetua a consulta utilizando as técnicas de extração do texto para expandir a pesquisa e detectar expressões negativas para então enviar a consulta ao módulo de recuperação da informação. Esse módulo recupera o conhecimento, classifica e envia para o usuário uma lista de documentos recuperados em ordem de relevância.

Para melhor funcionamento do modelo, ele foi divido em dois processos: o processo de indexação, onde a informação, que já está armazenada no banco de dados, é processada a fim de criar a base de conhecimento e o processo de recuperação, onde o usuário efetua das buscas propriamente dita.

5.4.1 Indexação do conhecimento

A Figura 18 apresenta uma visão de como o termo é indexado e como os usuários de recebem os resultados. Nesse modelo, as informações de textos vindos da base de laudos do STT, da base de toxicologia clinica do HU, ou da base de substâncias Perigosas (HSDB), são processadas pelo módulo "Pré-Processador". O resultado deste pré-processamento da informação é validado pelo especialista de domínio médico e então armazenado na base de conhecimento.

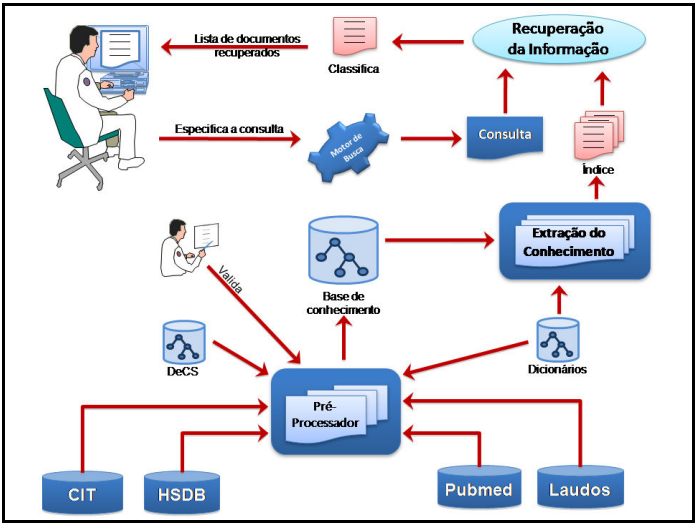


Figura 18: Modelo esquemático do sistema de recuperação de informação semântica a partir de bases de dados médicas.

A indexação dos termos que estão na base de conhecimento é efetuada pelo módulo de extração do conhecimento. Esse módulo é responsável por acessar a base de conhecimento e os módulos de extração de texto (Detecção de expressões negativas e expansão de pesquisa) e gerar um índice invertido para que a ferramenta de recuperação da informação possa extrair o conhecimento (Cabral, 2010).

Dentro do módulo de pré-processamento é efetuada a anotação semântica dos documentos. Nesse módulo, um processo chamado de *Analyzer*, efetua a conversão dos textos em termos. Os termos são usados para determinar quais os documentos que correspondem a uma consulta durante a pesquisa.

O *Analyzer* é o componente do processo de análise, que realiza uma série de operações para facilitar a indexação. Ele converte letras minúsculas em minúsculas, remove caracteres sublinhados, remove palavras comuns, tais como artigos e pronomes (*stop words*), extrai a raiz das palavras (processo chamado de *stemming*) anota semanticamente a sentença e efetua a detecção de frases negativas.

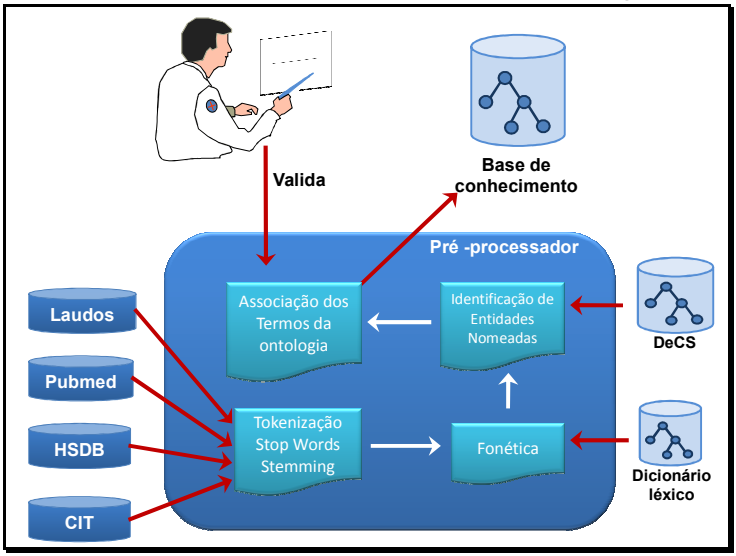


Figura 19: Módulo Pré-processador semântico

Após a normalização do texto, o módulo de fonética verifica se existe alguma inconsistência na informação recebida do banco de dados e atualiza a informação baseando-se no dicionário léxico da língua portuguesa. O analisador também é utilizado para extrair os termos da

ontologia DeCS, a fim de identificar as entidades reconhecidas na ontologia e associá-los com os termos processados. O especialista de domínio então valida as informações associadas pelo módulo de pré-processamento e este armazena o conhecimento na nova base. O módulo de pré-processamento pode ser visualizado na Figura 19.

No modelo proposto, é considerado que a base de conhecimento deve ser construída e associada a fontes de informação que utilizam ontologias médicas, dicionários léxicos e o *Analyzer* para indexar os conceitos que aparecem nos documentos armazenados no banco de dados médicos.

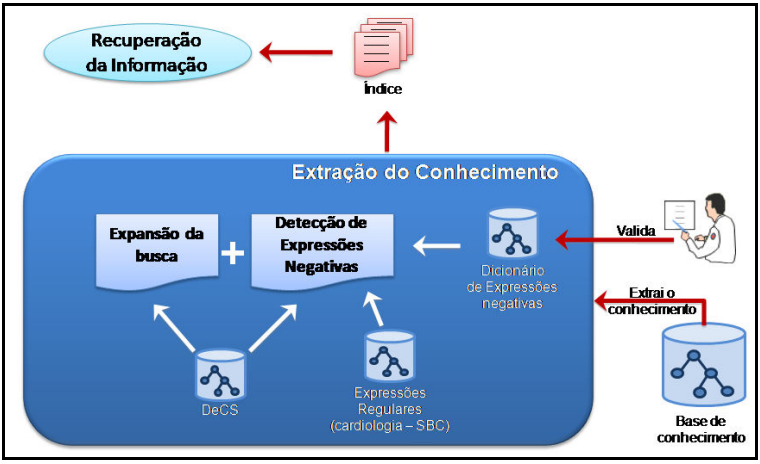


Figura 20: Módulo de Extração do Conhecimento

A Figura 20 demonstra como funciona o módulo de extração do conhecimento, que recupera a informação da base de conhecimento e envia a resposta ao módulo de recuperação. Ao recuperar a informação da base de conhecimento, esse conhecimento é direcionado ao método de expansão do universo do conhecimento, que verifica se o termo consta na ontologia DeCS e amplia o termo considerando os termos sinônimos e os termos relacionados.

Com o objetivo de melhorar ainda mais a precisão, o módulo de extração de termos negados é adicionado ao método de expansão. Para a confecção desse módulo, foi desenvolvido primeiramente um dicionário de expressões negativas mais utilizadas pela equipe médica. Esse dicionário foi criado a partir de um *Parcer* que analisou todos os termos da base do STT e do CIT e após foram validados pelos profissionais médicos. Ainda, como havia muitos erros de codificação, foi criado também um pequeno dicionário de expressões regulares para reparar

esses erros. Como resultado desse processo, é gerado um índice invertido com toda a informação processada.

5.4.2 Recuperação do conhecimento

O componente *de recuperação da informação* é responsável por conectar os módulos de extração, reformular a consulta e armazenar os resultados nessa base de índices invertidos. Os módulos de anotação semântica, expansão de pesquisa e detecção de expressões negativas são usados para processar os "*top-k*" resultados recuperados, ampliar a pesquisa e ranquear a informação ao usuário. O método de ranking é uma adaptação do modelo de espaço vetorial (Van Rijsbergen, 1975), que define pesos para os termos encontrados no documento. Esses pesos são computados automaticamente baseados na freqüência de instâncias em cada documento. O número de ocorrências para cada instancia do documento é definido pelo número de vezes que essa instancia aparece no texto.

A Figura 21 apresenta o modelo esquemático do funcionamento do módulo de pesquisa da presente proposta.

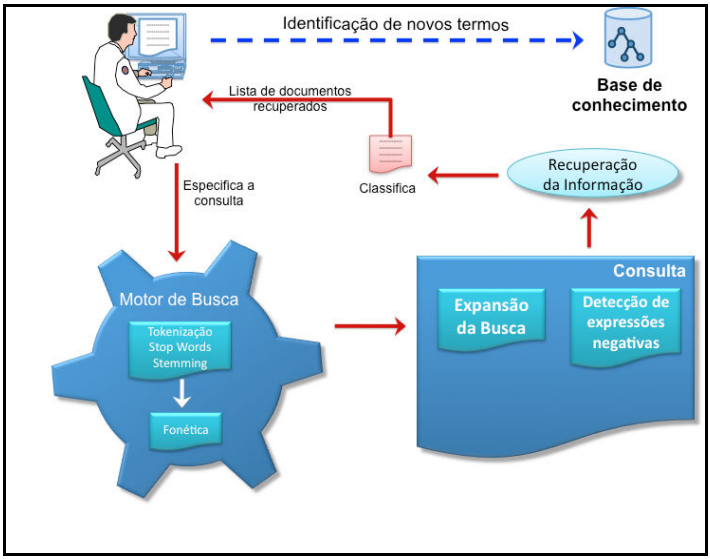


Figura 21: Módulo de pesquisa.

Quando um usuário efetua uma consulta, o motor de busca refina a pesquisa do usuário enviando os termos solicitados ao módulo de pré-processamento que *tokeniza* a expressão, elimina as *stop words*, e efetua o processo de *Stemming* da expressão solicitada pelo usuário. O resultado é enviado ao módulo de consulta que efetua a expansão da busca e a detecção das expressões negativas, para então enviar a nova pesquisa para o componente de recuperação da informação. Esse, por sua vez recupera os dados do índice invertido, classifica e envia uma lista com os documentos recuperados para o usuário.

5.5 AQUISIÇÃO E REPRESENTAÇÃO DO CONHECIMENTO

Como apresentado anteriormente, o uso e a disseminação do conhecimento são caracterizados por um conjunto de processos que tem como finalidade, auxiliar as organizações a gerenciarem seus conhecimentos. Dentro da área de saúde, essa disseminação do conhecimento pode ser representada por sistemas computacionais que permitem adquirir, criar, compartilhar e aplicar o conhecimento. Esse conhecimento é resultado de uma série de transformações que vão desde o armazenamento dos dados sobre uma realidade, até a interpretação das informações a fim de se obter uma ação. A capacidade de adquirir o conhecimento consiste na extração do conhecimento de um especialista ou a partir de bibliografias confiáveis e transpor para um sistema computacional com a finalidade de torná-lo inteligente.

Esse trabalho propõe o uso da ontologia DeCS para representar o conhecimento, no sentido de facilitar o reuso da informação armazenada nas bases de laudos médicos do STT da UFSC, na base de informações toxicológicas e de outras aplicações que possam ser incorporadas ao domínio. Nesse sentido, a principal razão para armazenar os documentos em bases de conhecimento é necessária para posterior recuperação e utilização. Para melhorar a indexação de documentos da área da saúde são utilizadas as ontologias de domínio médico, como DeCS, SNOMED, ou UMLS. Entretanto, se um documento não estiver referenciado por nenhuma ontologia, o processo de indexação e recuperação será dificultado, uma vez que o usuário poderá despender grande parte do seu tempo para filtrar as informações.

Ao analisar os textos armazenados na base do STT e na base de toxicologia do CIT, constatou-se que eles não foram indexados aos

termos das ontologias. Como conseqüência, uma pesquisa por determinado termo pode não ser encontrado na base de laudos. Por exemplo, se um usuário médico pesquisar na base de dados pela expressão "Fibrilação Auricular" a *engine* de pesquisa não retornará nenhum documento. Por outro lado, se o mesmo usuário pesquisar pelo termo "Fibrilação atrial", a *engine* de pesquisa retornará 4.260 laudos que correspondem à solicitação do usuário. Apesar dos termos serem sintaticamente diferentes, a ontologia DeCS define esses termos como sinônimos, nesse caso, são considerados semanticamente iguais. Se os laudos tivessem sido indexados por meio da ontologia, o usuário teria maiores informações sobre o assunto pesquisado e principalmente, não perderia tempo em ter que refazer sua pesquisa para encontrar a informação solicitada.

A aquisição do conhecimento por parte do sistema é feita em dois momentos distintos. No primeiro momento, os dados são analisados e extraídos pelo pré-processador que gerará a base de conhecimento. O segundo momento, de aquisição do conhecimento, é quando um usuário efetua uma busca contendo termos que são utilizados no seu cotidiano e que não estão armazenados na base de conhecimento. Nesse caso, essas informações serão armazenadas em uma base temporária que antes de ser efetivamente publicada para acesso publico, deve ser validada pelo especialista de domínio e incorporada à base de conhecimento.

As seções seguintes descrevem o funcionamento do modelo proposto, apresentando uma descrição detalhada do desenvolvimento do sistema, um pseudocódigo em alto nível e também a organização conceitual da base de conhecimento criada.

5.5.1 Normalização do texto

O primeiro passo para o desenvolvimento desse trabalho, foi a normalização do texto contido na base de dados, para somente depois passar para o processo de extração do conhecimento. Essa normalização se fez necessária, pois os laudos não continham somente elementos em texto livre, mas também elementos em linguagem HTML. Dessa forma, houve a necessidade de remover esses elementos HTML e alguns erros de codificação, como por exemplo, caracteres acentuados como "ã", quando interpretado em latin-1, transforma-se em "Ã£" no padrão UTF-8. Os elementos HTML, como quebra de linha (</br>), parágrafos (<p>) tamanho de letras (), dentre outros também foram removidos das

sentenças. Ao final constatou-se que em alguns laudos de ECG não havia ponto final de frase. Nesses casos, foi mantida a tag HTML </br> para sinalizar o final de uma sentença. A Figura 22 mostra um trecho da lista de expressões regulares encontradas na base de laudos de ECG do sistema de Telemedicina da UFSC. Essa lista foi composta após o processo de normalização.

```
ativa[çc][ãa]o\s+atrial\s+retr[óo]grada¶
TAM¶
supra\s*desn[ii]vel\s+[Ss][Tt]¶
infarto\s+atrial¶
infarto\s*(antigo|agudo((\s+do)?\s+mioc[áa]rdio)?)(\s+em\s+evolu[çc][ãa]o)?(\s+(de|em|na)
|septal|anterior|[âa]ntero[-·]?septal))?(\s+(e\s+)?(alta|dorsal))?¶
fibrila[çc][ãa]o\s+atrial¶
ritmo\s+(\w+\s+)?[Ff][Aa]¶
isquemia\s+(subepic[áa]rdic?a|subendoc[áa]rdic?a|sub(endo)?|mioc[áa]rdic?a|[âa]ntero)?¶
onda\s+T(\s+(invertida|negativa|apiculada|bif[áa]sica|isolada))?¶
s[ii]ndrome\s+(de\s+)?(brugada|wolff?(\s+|\s*-\s*)parkinson(\s+|\s*-\s*)white|pr[ée][-·
]excita[çc][ãa]o(\s+\(W\.P\.W\)|\s+ventricular\(W\.P\.W\))?)¶
W\.?P\.?W¶
fl[üu]tt?er(\s+atrial)?¶
taquicardia\s+(atrial(\s+parox[ii]stica)?|sinusal|juncional|parox[ii]stica\s+supra[-·
]?ventricular(supra[-·]?ventricular)?(\s+paroxistica)?(supra[-·]?ventricular)?(\s+n[ãa]o[-
TVNS¶
TPSV¶
ritmo\s+sinusal(\s+(taquic[áa]rdico|bradic[áa]rdic?o|juncional|idioventricular))¶
taquiarritmia(\s+(atrial|supra[-·]?ventricular))?¶
necrose¶
espa[çc]o\s+PR(\s+(curto))?¶
FA¶
rep(olariza[çc][ãa]o)?\s+vent(tricular)?¶
AD?RV¶
fibrose\s+((parede\s+)septal|dorsal|inferior|anterior|([âa]ntero[-·    ]?)?septal|lateral
]?dorsal)¶
BRA¶
BS¶
brad(i?cardia)?\s+sinusal¶
arritmia\s+sinusal¶
bloqueio\s+(do\s+)?ramo\s+(esquerdo|direito)¶
BC?RD¶
BC?RE¶
bloqueio\s+(divisional\s+([âa]ntero|p[óo]stero)[-·
]?(superior|inferior)\s+)?((in)?completo\s+)?(d[eo]\s+)?ramo\s+(direito|esquerdo)¶
bloqueio\s+([áàa]trio[-·]?ventricular|AV)¶
BDAS¶
```

Figura 22: Lista de expressões regulares encontradas nos laudos de ECG

Uma descrição do algoritmo de normalização do texto é descrito na Figura 23. Basicamente, ao receber um texto, o sistema percorre toda a base de laudos e para cada laudo, remove as *tags* HTML e corrige os erros de codificação.

```
1   Algoritmo Normalizar (laudo)
2   Entrada: laudo (texto)
3   Saída: texto normalizado
4   Inicio
5       Recebe laudo (texto)
6       Enquanto tiver laudo
7           Pesquisa tags HTML
8           Se encontrar
9               Remove as tags
10          Se encontrar erro de codificação
11              Remove erros a partir da lista de
expressões
12      Fim enquanto
13  Fim
```

Figura 23: Algoritmo de normalização de texto.

5.5.2 Expansão de pesquisas

O processo de expansão de pesquisa usa a ontologia DeCS adicionando mais informações médicas à pesquisa do usuário. A ontologia DeCS é utilizada para indexar o texto do relatório médico no banco de dados, pois esta ontologia contém conceitos, relações de sinonímia e conceitos relacionados. Isso facilita a expansão e a extração de termos dos descritores DeCS. Se o conjunto de expressões encontradas em uma pesquisa, contêm uma relação com a consulta do usuário e com os descritores do DeCS, será gerada uma nova consulta que contenham todos os termos presentes aos sinônimos do DeCS. Além disso, será usado um dicionário léxico para encontrar novas expressões sinônimas em relatórios médicos.

Por exemplo, ao se procurar pela doença "asma", diretamente na ontologia DeCS são encontrados quatro termos de resposta: "Asma", "Asma Induzida por Exercício" "Dispnéia Paroxística" e "Asma induzida por aspirina". Ao navegar pela hierarquia do DeCS, pode-se perceber que os três primeiros itens encontrados na pesquisa estão diretamente relacionados com a hierarquia do termo “asma” (Figura 24). Mas isso não quer dizer que a ferramenta de busca encontrou esse relacionamento. A busca foi feita utilizando a palavra chave somente, onde foram encontrados os termos que continham a pesquisa solicitada.

Figura 24: Fragmento da hierarquia do DeCS.

Baseado nos fundamentos do trabalho de Díaz-Galiano *et al.*, (2009), esse algoritmo foi adaptado para ser utilizado com a ontologia DeCS na língua portuguesa. O mecanismo de expansão da pesquisa analisa as entradas do usuário e procura na ontologia. Se for o termo da pesquisa for encontrado na ontologia, o algoritmo verifica quais os termos sinônimos, os termos relacionados e os descendentes em primeiro nível para expandir a pesquisa. Como resultado, a pesquisa expandida será a seguinte:

> "Asma", "Asma Brônquica", "Antiasmáticos", "Asma Induzida por Exercício", "Asma induzida por aspirina" e "Estado Asmático".

Ainda, o algoritmo permite usar o dicionário léxico que possibilita expandir os resultados incluindo três novos termos: "bronquite", "o vírus da bronquite infecciosa" e "Bronquite Crônica". Isso acontece porque, no dicionário usado, o termo "asma" tem relação com o termo "bronquite", mas na ontologia DeCS esta ligação não existe na mesma árvore hierárquica. A Figura 25 apresenta um exemplo da pesquisa inicial e o resultado do algoritmo de pesquisa expandida após o pré-processamento da informação.

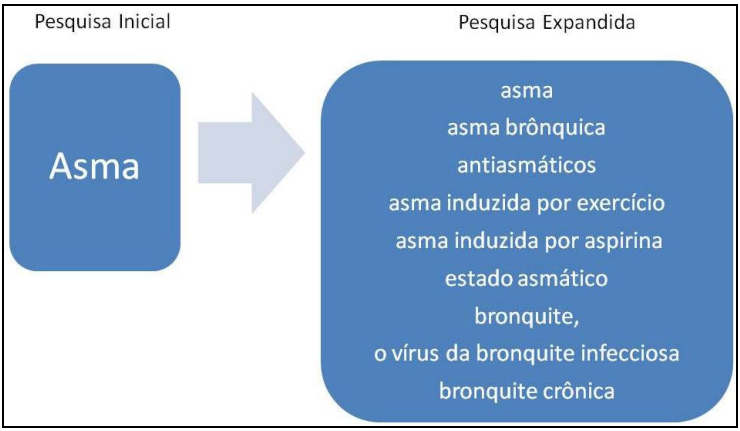


Figura 25: Exemplo de expansão de pesquisas usando a metodologia proposta.

Como a expansão da pesquisa pode conter informações que não foram incluídas na consulta do usuário, mas que possui alguma relação com a pesquisa inicial, foi necessário a definição de diferentes pesos semânticos nos termos da ontologia. A fim de representar uma relação semântica entre os termos, foram definidos pesos que variam de 0.9 (mais forte) até 0.5 (mais fraco). O Quadro 2 apresenta os pesos que foram associados às tipos de relacionamentos da ontologia.

Relacionamento	Peso
Termos sinônimos – DeCS	0.9
Termos relacionados - DeCS	0.8
Dicionário léxico – língua portuguesa	0.6
Descendentes em primeiro nível - DeCS	0.5

Quadro 2: Tipos de relacionamentos da ontologia e seus pesos semânticos.

Termos sinônimos possuem um peso maior que temos relacionados. Considerando o exemplo citado do termo "Asma" da ontologia que está na categoria "doenças", possui como sinônimo "Asma Brônquica" que está na mesma categoria (doença), mas possui como termo relacionado, "Antiasmáticos", que está classificado como "Compostos químicos e drogas". Apesar de terem relacionamentos semânticos, uma pesquisa por doença deve considerar esse termo menos relevante que as doenças.

O mesmo acontece para os outros dois relacionamentos. Os termos sinônimos encontrados no dicionário léxico não representam

necessariamente a mesma sinonímia que as que contam na ontologia, mas permitem aumentar a precisão dos resultados. No dicionário léxico utilizado, a doença “Bronquite” é sinônima de “asma”.

Apesar de bronquite ser uma doença que está classificada na mesma categoria da Asma (Doenças respiratórias), elas estão em subcategorias diferentes (Asma = C08.127.108 e Bronquite = C08.127.446). Com a definição desses pesos, os resultados apresentados pelo motor de busca têm pontuações (*scores*) diferentes e o próprio motor de busca já classifica o resultado em ordem descendente em função do *score* obtido.

Por exemplo, a Figura 26 apresenta um exemplo de uma consulta na base pelo termo Asma. A pesquisa foi propositalmente efetuada na base de dados da HSDB disponível na língua inglesa.

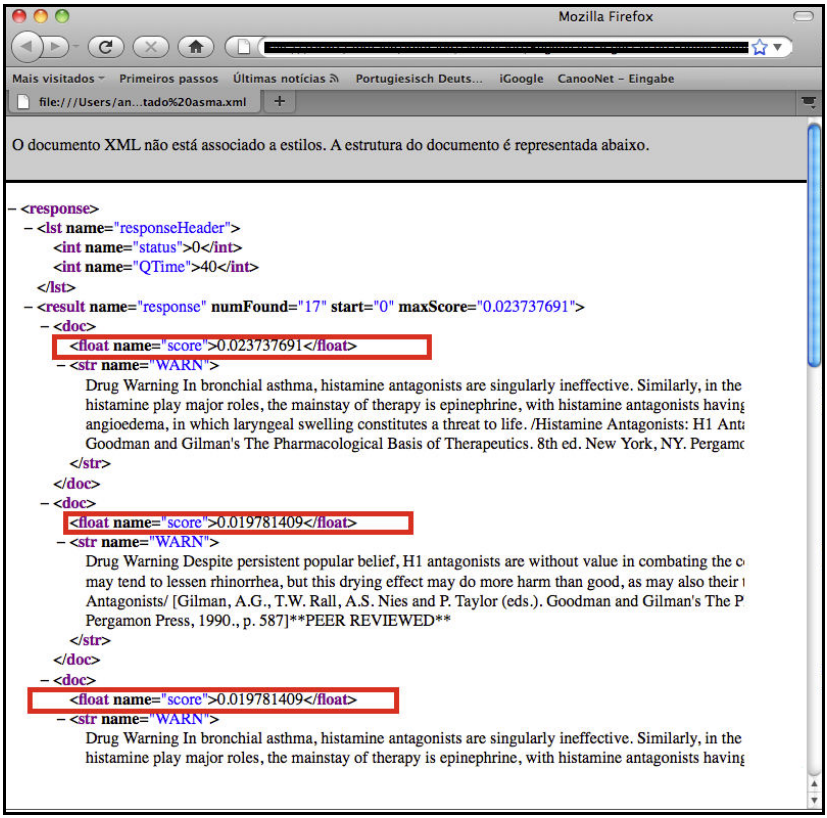


Figura 26: Exemplo de uma *query* com os pesos e *score* obtido

O termo foi pesquisado e encontrado no DeCS. Como existe uma descrição em três línguas para o termo, essa relação pode ser efetuada e retornada pelo mecanismo de busca. Mas como o termo exato não foi encontrado na base, o mecanismo de busca retornou os termos relevantes e de acordo com os pesos definidos no Quadro 2, foi montado o *score* obtido, classificado e enviado ao usuário as respostas já ordenadas.

Conforme descrito anteriormente, o método de expansão de pesquisa é responsável por analisar o texto do laudo, verificar se existem descritores que são identificados como idênticos na ontologia e a partir dessa identificação, encontrar os termos sinônimos, semelhantes e os termos que estão na mesma classe hierárquica da ontologia a fim de expandir os termos em uma pesquisa.

A descrição do algoritmo de expansão de pesquisa é descrito na Figura 27. Ao receber um termo de entrada, o algoritmo faz uma busca pelo termo para encontrar esse termo na ontologia DeCS. Se encontrar o termo, os termos sinônimos são armazenados em uma variável temporária, juntamente com a identificação única do descritor do DeCS, e é definido um peso de 0,9 para cada termo sinônimo ao termo definido como padrão. Em seguida o algoritmo verifica se existe algum termo relacionado a esse termo e define um peso de 0,8, caso encontre. O próximo passo é verificar a existência de termos na mesma hierarquia ou um nível abaixo, se encontrar, define peso 0,5 e armazena tudo nessa variável temporária.

O algoritmo, então, pesquisa se o termo procurado tem alguma relação com a lista de sinônimos da língua portuguesa (esse procedimento é verificado para suprir eventual inconsistência na ontologia, ou seja, se não houver um termo sinônimo na ontologia, mas na língua portuguesa, existir). Nesse caso, o algoritmo armazena o resultado na variável temporária e define um peso de 0,6 para o termo. Após, é verificado a existência de um termo de uso cotidiano pela comunidade médica. Se encontrar um termo cotidiano que tem referência com a ontologia, armazena na variável temporária o identificador da ontologia, o termo e o define peso de 0,9 para esse termo.

Como resultado, essa variável envia ao algoritmo de indexação para indexar a base de conhecimento e gerar os índices para as buscas.

```
1       Algoritmo Expansao (id, termo, posicao)
2       Entrada: Id, termo, posição no lauto
3       Saída: termo anotado
4       Inicio
5       Se encontrar termo DeCS
6               Resultado recebe (ID, termo, peso)
7               Para cada resultado
8                       Resultado recebe sinônimo (ID, Termo, peso = 0,9)
9                       Resultado recebe termo relacionado (ID, termo, peso =
0,8)
10                      Resultado recebe termo descendente (ID, termo, peso
= 0,5)
11      Pesquisar na lista de sinônimos LP
12              Resultado recebe sinônimoLP (ID, termo, peso = 0,6)
13      Pesquisar na lista de Termos cotidianos
14              Se encontrar resposta
15                      Resultado recebe termo cotidiano(ID, termo, peso =
0,9)
16      Fim
```

Figura 27: Algoritmo de expansão de pesquisa.

5.5.3 Detecção de expressões negativas

O modulo de detecção de expressões negativas é uma técnica de NLP que encontra sentenças negativas em resultados de exames médicos dentro da base de dados da telemedicina ou em outra base de dados médica. Como descrito por Mutalik *et al.*, (2001)(Mutalik *et al.*, 2001), será necessário detectar as classes negativas como negações adverbiais, preposicionais, adjetivas e verbos negativos para melhorar os parâmetros de precisão em uma *query*.

Conforme descrito anteriormente, o módulo de extração de termos negados foi desenvolvido a partir da base de laudos do STT, onde foi utilizada primeiramente uma base de dados de com 172 laudos de TC e 184 laudos de US. Após a fase de normalização, onde os laudos foram separados em sentenças, o algoritmo varre então toda a lista de laudos para encontrar as sentenças negadas.

O próximo passo foi encontrar as frases que continham termos negados. Para isso foi efetuado uma simples busca pelas expressões que continham uma única palavra negativa (visualizada na Figura 28). Como

resultado desse processamento, foi criado uma lista com os achados que continham uma palavra negativa.

ainda não
já não
Jamais
Nada
Não
não ... mais
Nem
nem mesmo
nem ao menos
nem ... sequer
nenhum(a)
Ninguém
Nunca
nunca mais
tão pouco
Negativo
Negativa

Figura 28: Lista de palavras com sentido negativo.

Como uma única palavra negativa não define que a expressão possa ser negada ou não, há a necessidade de analisar a expressão. Além disso, a maioria dos laudos médicos não dá 100% de certeza que a expressão encontrada é realmente negada ou não. Para isso foi criada uma lista com achados que foram chamados de "hipotéticos", como por exemplo, "Em D2 longo devido falha ECG **não se pode afirmar** extrassistolia supraventricular". Nesse caso, o profissional médico teve dificuldade em avaliar o traçado do ECG do paciente por não poder visualizar o sinal D2 que veio do equipamento de aquisição de imagem e não pode afirmar se o paciente possuía ou não algum achado clinico. Esse caso não deve ser interpretado pelo algoritmo como sendo uma expressão negativa. Dessa forma, foram definidas as expressões hipotéticas que serão avaliadas pelo especialista médico para validar a expressão. Outro caso, como, "Bloqueio divisional ventrículo-superior de ramo esquerdo. **Não se pode descartar** fibrose inferior". Nesse caso, o especialista médico não pode afirmar o achado clinico no exame do paciente baseando-se unicamente nas imagens dos sinais de ECG. Para

que o profissional pudesse avaliar com mais precisão, ele necessita de maiores informações, como por exemplo, indicação clinica, se o paciente toma algum medicamento, etc. No final desse processo, foi criada uma lista com as principais frases especulativas e/ou hipotéticas que foram encontradas na base de dados, conforme pode ser visualizado na Figura 29.

não sendo possível afastar
não se pode afirmar
não se pode descartar
não se pode descartar provável
não se pode descartar uma provável
não se pode excluir
não sendo possível afastar
porém não dá para descartar

Figura 29: Exemplo de uma lista de expressões hipotéticas.

O próximo passo no processo de detecção foi mapear os textos biomédicos encontrados na ontologia DeCS, para criar uma lista com os achados encontrados no DeCS e referenciar nos laudos do STT. Como resultado, foi gerado uma lista que foi indexada com os termos encontrados a fim de facilitar a pesquisa posterior. O Quadro 3 apresenta um trecho da lista de expressões negativas indexadas pelo algoritmo. A lista contém a identificação única (ID) do exame, a sentença que faz referência a esse ID, o termo negado e o(s) achado(s) encontrado(s) pelo algoritmo.

206876|Bloqueio divisional ântero-superior de ramo esquerdo, nao se pode excluir <achado>fibrose inferior</achado>

200330|Alterações da repolarização ventricular em parede inferior com onda T negativa em D3 e a Vf compatível com <achado>isquemia subepicárdica</achado>

295168|Não se pode descartar <achado>fibrose inferior</achado>

84079|<achado>taquicardia supraventricular não-sustentada</achado>

194868|Alterações difusas e inespecíficas da repolarização ventricular com ondas T negativas de V3 à V6 compatível com <achado>isquemia subepicárdica</achado>

42721|Não se pode descartar provável <achado>fibrose inferior</achado>

237727|Bloqueio divisional ântero-superior de ramo esquerdo , não se pode descartar <achado>fibrose inferior</achado>

93043|ARV septal, com onda T negativa em V1 e V2, compatível com <achado>isquemia subepicárdica</achado>

91614|onda T negativa em V1 e V2, compatível com <achado>isquemia subepicárdica</achado>
198830|Alterações da repolarização ventricular em parede septal não sendo possível afastar <achado>isquemia subepicárdica</achado>

100818|onda T negativa em parede septal, compatível com <achado>isquemia subepicárdica</achado>
91461|ARV com onda T negativa em V4 e V5, compatível com <achado>isquemia subepicárdica</achado>

101875|ARV com onda T negativa em D1, aVl, V3 e V4, compatível c/ <achado>isquemia subepicárdica</achado>

Quadro 3: Pequeno trecho da lista de frases negativas indexadas

Após os dados serem processados, o sistema identifica os termos encontrados no STT e compara com os termos encontrados na lista de diretrizes da Sociedade Brasileira de Cardiologia e para cada um dos termos identificados, o algoritmo classifica com negado ou não negado (nesse caso, o sistema também considera os termos hipotéticos). Como resultado, o sistema gera um documento XML anotado que será interpretado pelo módulo de extração do conhecimento (Andrade *et al.*, 2009).

O pseudocódigo disponível no algoritmo apresentado na Figura 30 mostra como um texto é anotado semanticamente pelo sistema. Basicamente o algoritmo recebe como entrada uma lista de termos disparadores que nada mais são que um conjunto de expressões que definem a possibilidade de um termo ser negado ou não. Esses disparadores são identificados como chave para a definição de uma expressão negativa. Os disparadores são definidos como:

- **pós-possibilidade**: existe a possibilidade de um termo negativo após a presença de um achado;
- **pré-possibilidade**: existe a possibilidade de um termo negativo antes a presença de um achado;

- **pós-negação**: a negação está presente após a definição de um achado e
- **pré-negação**: a negação está presente antes da definição de um achado.

O algoritmo recebe também o texto do laudo como entrada e, como saída, o laudo divido em sentenças anotadas com os termos negados e hipotéticos. Após a normalização do texto, o algoritmo separa o laudo em sentenças (um *tokenizador* de sentenças divide o laudo em varias sentenças) e, para cada sentença, os achados são identificados com base na ontologia e anotados semanticamente. A função "Anotar (sentença)" é descrita na Figura 30.

```
1   Algoritimo Negacao
2   Entrada: (termos_disparadores, laudo)
3   Saída: laudo dividido em sentenças anotadas com termos negados e
        hipotéticos
4   Inicio
5   Normalizar (laudo)
6   separar em sentenças (laudo)
7   Para cada sentença:
8           Identificar os achados na sentença (Laudo, DeCS)
9           Anotar (sentença)
10  Fim
```

Figura 30: Algoritmo para detecção de expressões negativas

Na Figura 30, linha 9, é apresentado a chamada do algoritmo que anota a sentença negativa. Basicamente esse procedimento descreve como identificar os termos disparadores e classificar os termos afirmativos que serão adicionados à lista de achados no laudo. Esse algoritmo recebe como entrada a lista de sentenças em cada laudo e envia como saída essas sentenças anotadas.

O primeiro passo é identificar as frases disparadoras na sentença e armazenar na lista de frases disparadoras. Após a definição dessas frases, o método *tokeniza* todas as informações (achados, frases disparadoras e as palavras que não foram identificadas no texto). Para cada *token* da sentença e se esse *token* é uma frase disparadora de **pré-negação** ou **pré-possibilidade**, o algoritmo define o escopo da frase disparadora, adicionando *tokens* à frente até que encontre o final de uma sentença. Se um termo de conjunção que define o final de uma sentença, ou se a frase disparadora é uma **negação** ou um **pseudo-negação,** for

encontrado, o algoritmo adiciona o termo encontrado na lista de frases disparadoras.

```
1     Algoritmo anotar (sentença)
2     Entrada: sentença
3     Saída: sentença anotada
4     Inicio
5     lista_frases_disparadoras <- identificar as frases disparadoras na
      sentença
6     tokenizar os achados
7     tokenizar as frases disparadoras
8     tokenizar as palavras que restaram
9     Para cada token da sentença:
10        se token é uma frase disparadora de pré-negação ou pré-
      possibilidade:
11    definir escopo da frase disparadora, adicionando tokens à frente até
      que encontre uma das seguintes condições:
12    *fim da sentença
13            *termo de terminação
14            *frase disparadora de negação ou pseudo-negação
15            adicionar a lista_frases_disparadoras
16        senão se token é uma frase disparadora de pós-negação ou pós-
      possibilidade:
17    definir escopo da frase disparadora, adicionando tokens antecedentes
      até que encontre uma das seguintes condições:
18            *início da sentença
19            *6 tokens já foram adicionados
20            *termo de terminação
21            adicionar a lista_frases_disparadoras
22        senão se token é um achado:
23            classificar o achado como 'afirmativo' por padrão
24            adicionar a lista_achados
25            Para cada frase em lista_frases_disparadoras:
26                    Para cada achado em lista_achados:
27                      se achado está no escopo da frase:
28                        se frase é pré-negação ou pós-negação:
29                              classificar achado como 'negativo'
30                        senão se frase é pré-possibilidade ou pós-
      possibilidade:
31                    classificar achado como 'especulativo'
32            Fim para
33    Fim
```

Figura 31: Algoritmo para o procedimento anotar sentença

Caso o *token* seja uma frase disparadora de **pós negação**, ou **pós-possibilidade**, o algoritmo define o escopo da frase disparadora, adicionando *tokens* antecedentes até que encontre o inicio de uma sentença, ou um termo de conjunção que define o final da sentença ou então, se seis *tokens* já foram adicionados à sentença. Nesse caso, ele adiciona o termo encontrado na lista de frases disparadoras. E caso o *token* não for um achado, ele classifica o achado como "afirmativo" e adiciona a lista de achados.

Depois que a sentença é anotada, o próximo passo é o ordenar a lista de frases disparadoras (Figura 31, linha 25) pela ordem de precedência. A ordem ascendente é definida pelas frases que contem os disparadores na seguinte seqüência: pós-possibilidade, pré-possibilidade, pós-negação e pré-negação. Se o achado encontrado na lista do DeCS for classificado como pré-negação ou pós-negação e estiver no escopo da frase, o algoritmo classifica esse achado como "negativo", caso contrário (o achado for classificado como pré-possibilidade ou pós-possibilidade), ele classifica esse achado como expressão "especulativa" (nesse caso, a sentença será considerada como sendo hipotética).

5.5.4 Anotação semântica

Para definir a anotação semântica, foi usado um dicionário léxico da língua portuguesa, um dicionário de frases negativas e hipotéticas e uma lista de termos utilizados pelos profissionais médicos no seu dia-a-dia (lista de termos cotidianos). Para a implementação da base de conhecimento foi necessário criar uma modelagem de banco de dados para armazenar as anotações semânticas devido à característica não-intrusiva do algoritmo a ser desenvolvido. Ou seja, todas as anotações semânticas reconhecidas pelo algoritmo não foram armazenadas nos laudos.

Além disso, a ontologia DeCS foi utilizada para criar anotação semântica de texto automaticamente e armazenar na base de conhecimento. O objetivo dessa base não é apenas recuperação termos constantes nas consultas, mas também a recuperação de entidades relacionadas, sinônimos e termos relacionados armazenados hierarquicamente na ontologia médica. Usando semelhanças estruturais do documento, foram identificadas entidades nomeadas, associadas a estas entidades com conceitos, validadas por um especialista médico e

definidas as anotações semânticas. Ainda, para os termos que são utilizados pelos médicos em seu dia-a-dia são armazenadas em uma base de dados após a validação do especialista. Essa validação é efetuada referenciando o termo de uso cotidiano à ontologia DeCS, quando houver alguma referência. A Figura 32 mostra um pseudocódigo do algoritmo de anotação semântica utilizado no desenvolvimento dessa tese. Basicamente, o algoritmo para cada termo da lista da ontologia, ele percorre a base de laudos em busca do mesmo termo.

```
1   Algoritmo anotaçãoSemantica (laudo)
2   Entrada: Laudo
3   Saída: sentença anotada
4   Inicio
5   Decs recebe lista de termo do decs
6   Laudo recebe lista de laudos medicos
7   Anotaçoes recebe lista de anotações semanticas
8   Para cada termo em decs
9       Para cada laudo em laudos
10          Para cada sentença em laudo
11              Se sentença contem termo
12              Adiciona
                anotacoes(sentença,termo,posicao,tamanho)
13  Fim
```

Figura 32: Algoritmo de anotação semântica.

Nesse caso, o algoritmo de anotação semântica é relativamente simples. O algoritmo tem como entrada os laudos que estão armazenados no banco de dados e como saída, o algoritmo envia para o analisador o laudo anotado semanticamente. O processo inicia com as variáveis recebendo a lista de termos do DeCS, a lista de laudos e uma lista das anotações semânticas, se houver alguma já anotada. Na Figura 32, linha 8 o algoritmo inicia um loop que corresponde à cada termos encontrado no DeCS. Para cada sentença encontrada nos laudos e em cada sentença contém um termo disparador (nesse caso um achado), o algoritmo adiciona as anotações na sentença. Ainda, ele armazena a posição onde foi encontrado o termo e qual o tamanho desse termo em quantidade de caracteres. Ao final, tem-se a lista anotada com todos os termos da ontologia referenciados aos termos da base de laudos.

5.5.5 Analisador

Conforme apresentado na seção anterior, antes de indexar um termo, há a necessidade de pré-processar um documento. Essa fase de pré-processamento é chamada de analisador, ou (*Analyzer*). Esse analisador é responsável por eliminar as *stop words,* efetuar a *tokenização,* definir o *stemming* para o termo, quando necessário, identificar as entidades nomeadas da ontologia e associar os termos da ontologia aos termos do banco de dados. Para o desenvolvimento desse trabalho, foi utilizado e adaptado a classe *BrazilianAnalyzer* do *Lucene* para que atendesse às necessidades do projeto. Esse analisador é um filtro para a efetiva entrada de dados e serve como pré-processamento antes enviar os dados parada classe *IndexWriter*. Uma descrição em alto nível do algoritmo analisador é apresentada na Figura 33.

```
1       Algoritmo analisador (lista de frases do laudo)
2       Entrada: laudo normalizado (texto)
3       Saída: Laudo dividido em sentenças anotadas (texto)
4       Inicio
5               Recebe (laudo normalizado)
6               Mapeia Stop words a partir da lista para a língua portuguesa
7               Separar em sentenças (Laudo)
8               Para cada senteça ϵ laudo Faça
9                       Remove Stop words
10                      Converte texto para caixa baixa
11                      Reduz ao radical da palavra (chama Classe BrazilianStemmer)
12                      AnotaçãoSemantica (laudo)
13                      Expande a pesquisa (ID, termo, posição)
14                      Detecta expressões negativas (ID, termos disp, laudo)
15              Fim Para
16      Fim
```

Figura 33: Algoritmo analisador do texto

O algoritmo analisador, descrito em pseudocódigo, foi alterado para basicamente chamar a classe de expansão da pesquisa descrito na seção anterior que a partir do mapeamento das sentenças que foram divididas e anotadas com a Ontologia DeCS, permitirá a indexação de

todo o conjunto de informações relacionadas. Esse algoritmo tem como entrada o laudo que foi normalizado pela classe que removeu as *tags* HTML e gerou um texto limpo. Como saída, o algoritmo envia para o indexador a lista de sentenças anotadas semanticamente.

O analisador inicia com o mapeamento das *stopwords* que são definidas na lista que está contida na classe *BrazilianAnalyzer* do *lucene* (Figura 33, linha 6). O algoritmo então percorre toda a lista de laudos encontrados na base de dados e para cada sentença pertencente ao laudo ele remove as *stopwords*, converte o texto para caixa baixa, reduz ao radical da palavra, expande a pesquisa e detecta se existem expressões negativas em uma sentença.

As linhas 12, 13 e 14 chamam as classes que foram descritas na seção anterior, que servirão respectivamente para anotar semanticamente um laudo, expandir a quantidade de termos sinônimos e detectar se existe alguma expressão ou sentença negativa no laudo. Como resultado tem-se o laudo pré-processado para que o algoritmo de indexação possa gerar o índice invertido.

5.6 INDEXAÇÃO DA BASE DE CONHECIMENTO

Conforme descrito anteriormente, a indexação do texto utilizado nesse trabalho de pesquisa é efetuado pela API do *Lucene IndexWriter()*. Essa API armazena as informações de entrada em uma estrutura de dados chamada de índice invertido. O índice possibilita aos usuários efetuarem buscas rápidas com a finalidade de localizarem os documentos correspondentes a uma determinada consulta. O índice invertido fornece métodos para incluir, excluir ou atualizar os documentos em um índice (Hatcher e Gospodnetic, 2004).

Após os dados serem processados pelo método *Analyzer()*, eles são incluídos no índice. Cada documento a ser indexado possui um identificador que contém as informações características sobre esse documento, tais como as ocorrências de uma palavra, a localização e a freqüência em que uma palavra aparece no texto. O algoritmo de indexação constrói uma lista de entradas, sendo que cada uma dessas entradas contem um identificador e o peso de cada termo. O índice gerado por essa tese possui várias entradas de dados que são apresentadas no modelo conceitual da base de conhecimento da Figura 34.

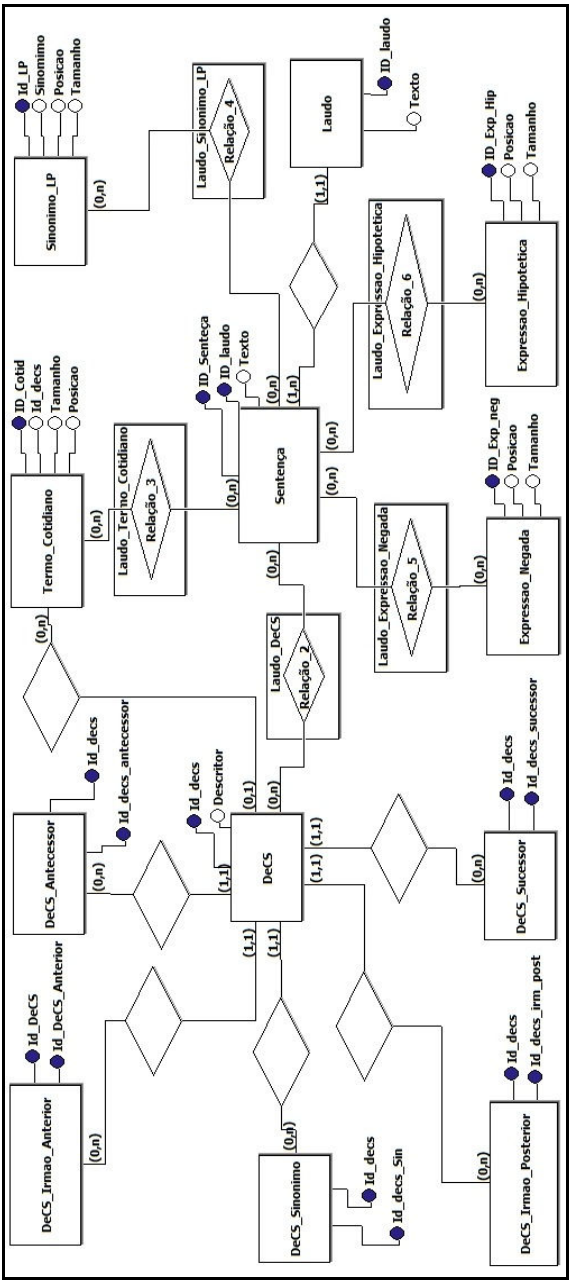


Figura 34: Modelo conceitual da base de conhecimento.

No modelo conceitual gerado, a base de conhecimento é descrita pelas entidades: laudo, DeCS, Termo_Cotidiano, Sinonimo_LP, Sentença, Expressao_negada e Expressão_Hipotetica. A entidade DeCs por sua vez é descrita pelas entidades DeCS_sinonimo, DeCS_Irmao_anterior, DeCS_Irmão posterior, DeCS_Antecessor e DeCS_Sucessor. Como a entidade *Laudo* e *Sentença* não podem ser relacionadas diretamente com as outras entidades, há a necessidade de criar associações por meio da criação de entidades associativas. Esse modelo conceitual consiste em criar dinamicamente uma visão para relacionar os dados contidos nas entidades e gerar o índice invertido.

Essa visão criada em forma de entidades associativas é aqui chamado de base de conhecimento, pois essa base reúne em uma única visão todas as tabelas necessárias para a aquisição e comunicação do conhecimento.

Da mesma forma, na Figura 35 é apresentada a mesma visão esquemática do modelo lógico da base de conhecimento. No modelo lógico aqui descrito, são apresentadas somente as chaves primárias para efetuar a associação das entidades. A partir da criação dessa visão lógica do modelo a base de conhecimento está relacionada com as outras tabelas no banco de dados e permite então a criação dos índices. O método que efetivamente faz a criação dos índices é chamado *indexWriter()* e usa essa visão lógica para criar os índices.

É importante frisar que o banco de dados é puramente um repositório de dados de onde as informações são recuperadas para gerar a base de conhecimento. O processamento dessas informações contidas na base é feito a partir da visão que foi gerada pelo repositório e processada pela API do *lucene*. O *Lucene* armazena os dados de entrada em uma estrutura de dados que é chamado de índice invertido. Esse índice é armazenado em sistema de arquivos (em um diretório do servidor), que a partir desse diretório serão processadas as consultas dos usuários.

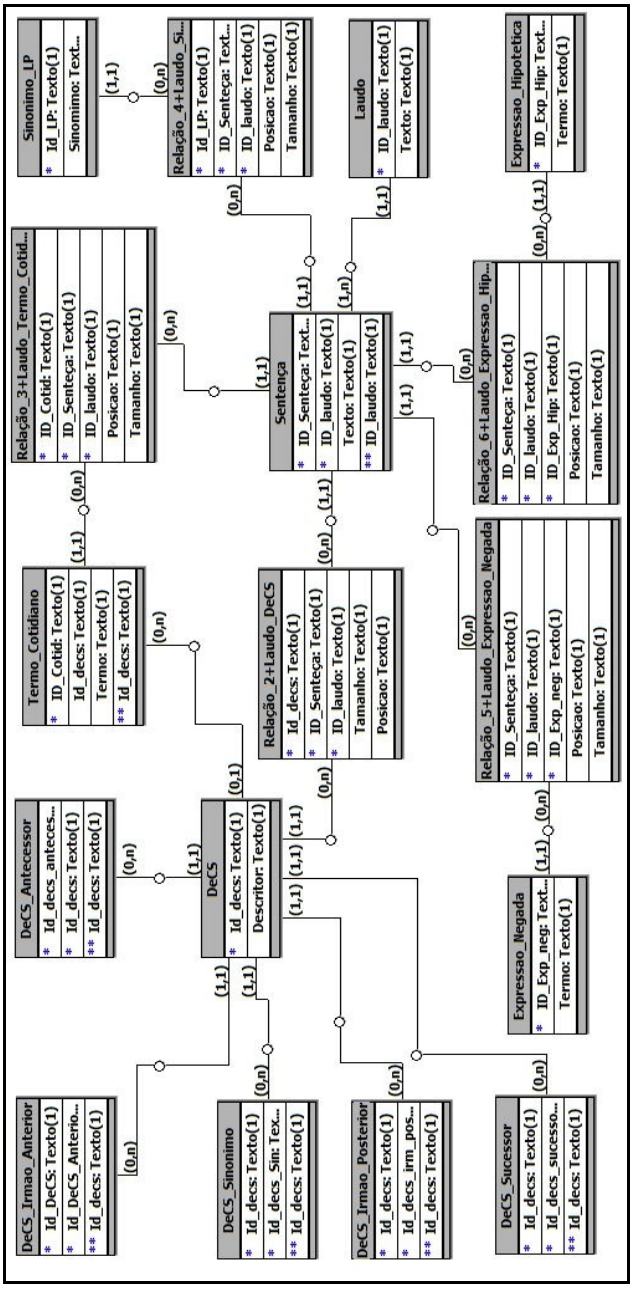


Figura 35: Modelo lógico da base de conhecimento.

A Figura 36 mostra o algoritmo para indexar as informações da base de conhecimento e gerar o índice invertido. O algoritmo recebe como entrada as tabelas contendo as informações da base de conhecimento e como saída, o índice invertido é criado.

```
1     Algoritmo IndexWriter()
2     Entrada: base de conhecimento
3     Saída: índice invertido
4     Inicio
5           Indexador Recebe o caminho do banco que contém os
            documentos a serem indexados;
6           Se existem documentos Então
7                 Cria um novo índice
8                 Repita para cada documento
9                       Criar entrada para palavra-chave no laudo;
10                      Criar entrada para Posição no laudo
11                      Criar entrada para termo dicionário_LP
12                      Criar entrada para termo DeCS
13                      Criar entrada para Expresao_Negativa
14                      Criar entrada para Expressao_Hipotetica
15                      Criar entrada para Termo_Cotidiano
16          Caso contrário
17                Retorna Mensagem "Não há documentos para
indexar"
18    Fim
```

Figura 36: Algoritmo para a criação do índice invertido

Em seguida, o algoritmo verifica se existem documentos na base de conhecimento, se houver, para cada documento encontrado ele sugere a criação de uma entrada no índice com as seguintes informações da base de conhecimento:

- Termos anotados semanticamente na ontologia;
- Posição de cada termo que foi anotado pela ontologia;
- Termos que foram encontrados no dicionário da língua portuguesa que estão presentes no laudo, mas não estão presentes na ontologia;
- Termos que contém expressões negativas;
- Termos que contém expressões hipotéticas;
- Termos que foram solicitados nas pesquisas anteriores e validados por especialista de domínio (termo_cotidiano).

Esse conhecimento é então indexado pelo algoritmo e permite que o mecanismo de busca possa efetuar as pesquisas muito rapidamente, sem que haja a necessidade de percorrer todas as bases de dados em busca da informação.

6 AVALIAÇÃO (APLICAÇÃO DO MODELO)

Para a avaliação do desempenho do protótipo, foram realizados experimentos baseados na abordagem GQM (*Goal Question Metrics*) (Basili *et al.*, 1994). Segundo o autor, o resultado dessa abordagem é a especificação de um conjunto de medidas que tem como objetivo, definir um conjunto de regras para interpretação dos dados. Esse modelo possui três níveis de mensuração: nível conceitual (Objetivo), nível operacional (Questões) e nível quantitativo (métricas). A Figura 37 mostra o modelo apresentado pela abordagem GQM para a avaliação da qualidade de software.

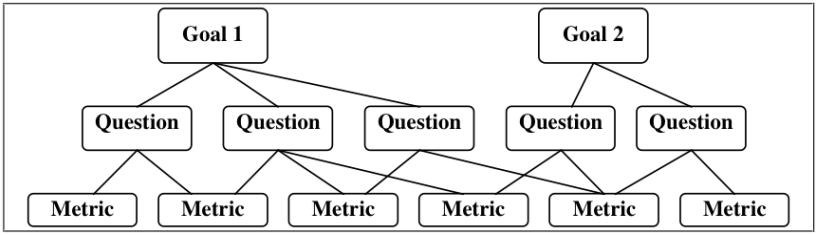


Figura 37: Modelo de uma abordagem do GQM (Basili *et al.*, 1994).

Ao utilizar a abordagem GQM, essa tese tem como objetivo verificar o comportamento do protótipo e a correspondência com o modelo teórico apresentado. Os objetivos desses experimentos são executados para avaliar a eficácia do modelo de recuperação e comunicação do conhecimento gerado pelo modelo no domínio da saúde.

Para a avaliação foram levadas em consideração as variáveis quantitativas e qualitativas, tais como, índice de precisão (quantidade de acertos), *recall*, tempo de resposta e a analise dos termos quando da extração do conhecimento. Dessa forma duas principais questões foram levantadas para a avaliação da performance:

1) Qual a proporção de resultados relevantes em relação aos resultados obtidos em cada consulta realizada no experimento?
2) Qual a proporção de resultados relevantes obtidos em relação ao total de resultados relevantes em cada experimento?

Para atender o modelo proposto pela abordagem GQM, as métricas utilizadas nesse estudo, foram *Precision* (precisão total), *Average Precision* (AveP), que representa uma média de precisão calculada sobre um conjunto de consultas (Buckley e Voorhees, 2000) e

a *Precision at Ten (P@10)*, que considera os dez primeiros registros para avaliar a precisão (Van Rijsbergen, 1975).

Para a definição da precisão, que representa a percentagem dos documentos recuperados que são relevantes para a consulta, foi utilizada a seguinte formula:

$$\text{Precision} = \frac{|A \cap R|}{|A|}$$

Onde: $A \cap B$ representa o total de documentos relevantes recuperados e A representa o total de documentos relevantes em uma consulta.

Para definir o cálculo do *Recall*, que representa a percentagem de todos os documentos relevantes para a consulta que foi recuperada, a formula utilizada foi a seguinte:

$$\text{Recall} = \frac{|A \cap R|}{|R|}$$

Onde: $A \cap B$ representa o total de documentos relevantes recuperados e R representa o total de documentos recuperados em uma consulta.

A métrica de *Recall* é utilizada em conjunto com a métrica de *precision* para calcular a *AveP*. A métrica *AveP* associa em uma mesma pontuação os valores clássicos *precision* e *recall*. Ela é utilizada para avaliar as consultas e é definida pela média de precisão computada no ponto de cada um dos documentos relevantes em uma seqüência classificada. É definida pela seguinte fórmula:

$$AveP = \frac{\sum_{r=1}^{N} (P(r) x\ rel(r))}{|A|}$$

Onde r representa o *rank*, N o número recuperado, *rel()* representa uma função binária sobre a relevância de um determinado *rank*, *P(r)* a precisão em um determinado recorte (relevantes documentos recuperados do *rank r* divido pelo *rank*).

A métrica P@10 é utilizada para definir a precisão de documentos recuperados quando o universo total de documentos relevantes não é conhecido. Essa métrica é muito utilizada em ferramentas de buscas para *web*, pois normalmente o usuário está

interessado somente nos documentos recuperados que são apresentados na primeira página do buscador da Internet. Quando o usuário não está satisfeito com a resposta, ele refina sua busca e novos resultados são apresentados pelo motor de busca. Essa métrica é definida pela seguinte fórmula:

$$P_k = \frac{Relevantes\ Recuperados\ em\ k}{k}$$

Onde P_k representa a proporção de documentos relevantes que foram recuperados dentre os primeiros *k* documentos recuperados (nesse modelo como são utilizados os 10 primeiros termos como métrica de avaliação, o *k* é igual a 10). *Relevantes Recuperados em k* representa o número de documentos relevantes dentre os *k* (10) documentos recuperados.

6.1 VALIDAÇÃO DA BASE DE CONHECIMENTO PELOS ESPECIALISTAS

Para validar uma base de conhecimento são necessárias pessoas que possuam um alto grau de conhecimento em determinado domínio e habilidade para transmitir o conhecimento. Essas pessoas são chamadas de especialistas de domínio. Os especialistas ajudam a estruturar o conhecimento e permitem minimizar os erros indexação errônea dos sistemas computacionais.

O objetivo da validação do especialista de domínio médico nessa tese é permitir a criação de uma base de conhecimento anotada, para que seja possível a representação e comunicação do conhecimento em pesquisas futuras. Dessa forma, serão anotadas as bases de toxicologia e alguns laudos da base do STT, com o objetivo de melhorar os resultados das pesquisas dos usuários. Ao final serão computados os resultados e serão efetuadas as métricas de precisão para validar a presente proposta.

6.1.1 Anotação da base toxicológica

Nesse estudo foram realizadas quatro validações de diferentes bases de dados por diferentes especialistas de domínio na área médica. No primeiro estudo foram realizadas diversas consultas com a

ferramenta de pesquisa na área de toxicologia clinica. Foram selecionadas 50 pesquisas por palavra-chave no domínio de classes de substâncias que estavam presentes na ontologia de toxicologia (Base CIT que aparece representada na Figura 19). Cada consulta foi selecionada aleatoriamente pelo próprio sistema, mas com o objetivo de garantir imparcialidade na avaliação e na escolha dos termos selecionados para o teste.

O objetivo desse experimento é avaliar se uma determinada substância é relevante quando um usuário efetuar uma pesquisa por um termo diferente, mas que tenha algum tipo de relacionamento, como por exemplo, sinônimo, mesma classe hierárquica na ontologia, entre outros. A avaliação foi feita por um especialista da área de toxicologia onde esse especialista efetuava a pesquisa e validava as respostas do sistema para cada um dos 50 termos selecionados anteriormente. Os resultados desse estudo serão apresentados na seção 7- Resultados Experimentais e Discussões.

6.1.2 Anotação da base de laudos

Para o desenvolvimento desse estudo foi implementada uma ferramenta simples para que os especialistas pudessem validar a base de conhecimento que foi criada. O desenvolvimento desse método de avaliação foi baseando nos estudos de Chapman (2001), onde o autor um conjunto de 1000 sentenças que seriam avaliadas por três especialistas da área e criar um "padrão ouro". A avaliação dos especialistas consistia em julgar as sentenças e definir uma das três respostas:

a) Presente – o termo destacado está presente na sentença;
b) Ausente – o termo destacados está ausente na sentença e
c) Ambíguo – o termo presente ou ausente não estava claro para o especialista na sentença.

Cada especialista médico julgou 400 das 1000 sentenças, sendo que 200 sentenças foram sobrepostas e avaliadas pelos 3 especialistas para determinar a melhor precisão das respostas. Se dois especialistas ao julgares a mesma frase não concordassem se o termo estava presente ou ausente, o termo era considerado como ambíguo.

Para adaptar à realidade da linguagem portuguesa médica, foi necessário acrescentar mais um item à lista, o termo hipotético. Esse termo é bastante utilizado na linguagem médica, pois o especialista de

posse somente de um exame de imagem não pode afirmar com 100% de certeza se um achado está ou não presente na imagem. Ainda, foi adicionada a expressão "termo em destaque não é um achado". Essa expressão é utilizada para que o especialista pudesse validar se o algoritmo de detecção de expressões negativas conseguiu anotar a expressão corretamente baseado no DeCS.

A Figura 38 mostra o exemplo da interface de validação do especialista em laudos de TC para analise das sentenças recuperadas pelo sistema. O objetivo dessa tela é agilizar o processo de validação por parte do especialista.

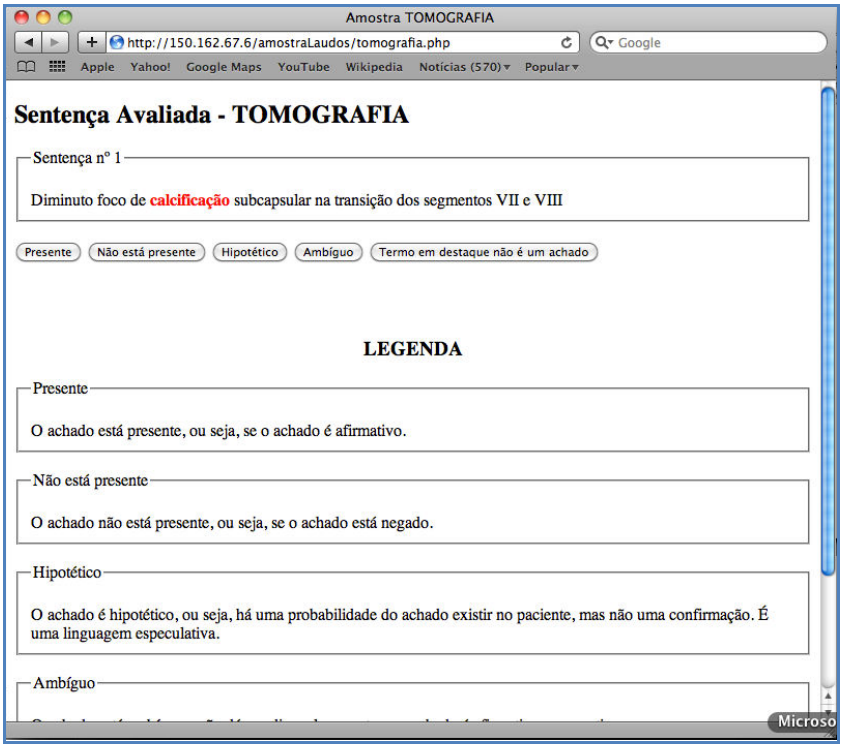


Figura 38: Interface para validação do especialista em TC.

O especialista recebe uma sentença que foi detectada pelo sistema e ele necessita validar essa sentença. Para isso, ele precisa selecionar uma opção das cinco que aparece para ele. São elas:

1. **Presente** – O achado está presente na sentença em destaque, ou seja, o achado é afirmativo;

2. **Não está presente** – O achado não está presente na sentença em destaque, ou seja, o achado é negativo
3. **Hipotético** - O achado pode ser considerado hipotético, ou seja, há uma probabilidade do achado existir no paciente, mas não se pode confirmar essa sentença;
4. **Ambíguo** - O achado pode ser considerado ambíguo. Não é possível afirmar claramente se o achado é afirmativo ou negativo e
5. **Termo em destaque não é um achado -** O termo em destaque não é um achado clínico.

Esse experimento foi validado em duas modalidades diferentes de exames, TC e US. Para anotar a base de conhecimento de TC, foram escolhidas aleatoriamente 172 sentenças de TC e 184 sentenças de US. Especialistas médicos do Hospital Universitário da UFSC/SC foram selecionados para validar essas sentenças.

Tabela 2: Resultado da avaliação do especialista em laudos de TC

Avaliação	**%**	**Qtde**
Presente	83%	142
Não está presente	11%	19
Hipotético	0%	0
Ambíguo	2%	4
Tema em destaque não é um achado	4%	7
Total	**100%**	**172**

A Tabela 2 mostra o resultado da avaliação do especialista de domínio médico da área de radiologia do HU/UFSC para os laudos de TC. O especialista encontrou 142 laudos que foram categorizados como afirmativo; 19 laudos que foram categorizados como negativos; quatro como ambíguos e sete laudos que não são achados (o sistema identificou erroneamente). Nesse caso, pode-se perceber que somente para esses 172 laudos, um sistema tradicional iria indexar esses 19 laudos (11%) como sendo afirmativo e as ferramentas de busca não saberiam definir se um laudo não deverá fazer parte da resposta ao usuário. A Figura 39 apresenta um gráfico com as percentagens dos resultados avaliados pelo especialista para laudos de TC.

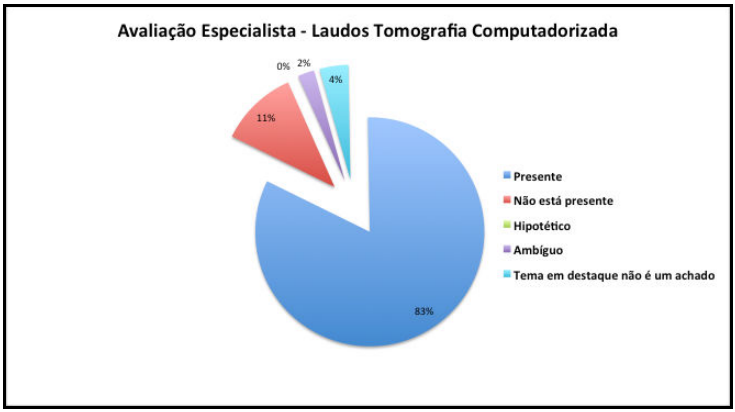


Figura 39: Resultado da validação dos laudos de TC.

O mesmo procedimento foi executado para laudos de US. A Tabela 3 apresenta o resultado da validação do especialista em laudos de US.

Tabela 3: Resultado da avaliação do especialista em laudos de US

Avaliação	**%**	**Qtde**
Presente	39%	71
Não está presente	29%	53
Hipotético	4%	7
Ambíguo	1%	2
Tema em destaque não é um achado	28%	51
Total	**100%**	**184**

Aqui se pode perceber que 51 laudos foram anotados erroneamente pelo sistema de anotação semântica. O especialista validou como não sendo um achado. Em compensação, essa base de US continha 53 sentenças que foram validadas como sendo expressões negativas.

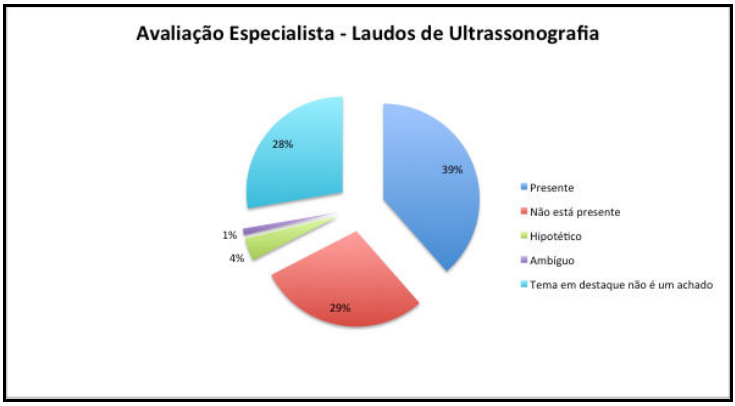


Figura 40: Resultado da validação dos laudos de US.

7 RESULTADOS EXPERIMENTAIS E DISCUSSÕES

Uma vez desenvolvido o protótipo, vários experimentos foram realizados para verificar o comportamento do sistema de busca em situações reais e nesse caso verificar a conformidade com o modelo teórico aqui descrito. Os trabalhos de testes foram elaborados no sentido de validar o processo de recuperação e comunicação do conhecimento no domínio da saúde, objetivo principal deste trabalho.

Os experimentos foram separados em três categorias para validar a pesquisa primeiramente com o desenvolvimento da expansão de pesquisa com base na ontologia. O segundo experimento apresenta os resultados da implementação da técnica de detecção de expressões negativas em comparação com as tradicionais técnicas de RI. E por fim, no terceiro experimento foram unidas as duas técnicas (expansão de pesquisa usando anotações semânticas juntamente com Detecção de textos negativos) para testar a qualidade das buscas.

O objetivo geral dos três experimentos é estudar o desempenho do mecanismo de busca no aspecto de utilização da semântica juntamente com modelos de recuperação de informações na área da saúde. O ambiente experimental foi projetado para ser o mais próximo possível das tarefas de um profissional de saúde. Para tanto, as pesquisas foram selecionadas aleatoriamente dentro do ambiente de informações toxicológicas e dentro de um ambiente de radiológico.

O primeiro teste consiste em escolher um determinado termo e um especialista de domínio, que percorre a lista inteira da base de conhecimento para encontrar termos sinônimos e semelhantes ao encontrado. Esse procedimento faz-se necessário, pois é preciso saber qual é a quantidade de possíveis resultados corretos que o sistema poderá retornar (recall). Essa é a fase de validação da base de conhecimento.

Num segundo momento um usuário qualquer efetua diversas pesquisas no sistema em busca de conhecimento. Os resultados serão confrontados com a base anotada para definir a quantidade de documentos que foram retornados corretamente (precisão).

7.1 PRIMEIRO EXPERIMENTO – EXPANSÃO DE PESQUISA

Para validar o método de expansão de pesquisas foram efetuadas várias pesquisas usando as métricas de precision e Recall. Nessa seção

serão comparadas as técnicas de RI utilizando expansão de pesquisa simples e com aprimoramento semântico, contra a tradicional técnica de RI. Os experimentos foram efetuados utilizados com base na base de conhecimento construída previamente por Junior et al., (2009). Em uma primeira fase, o especialista em toxicologia efetuou 50 pesquisas onde os resultados são apresentados na Figura 41.

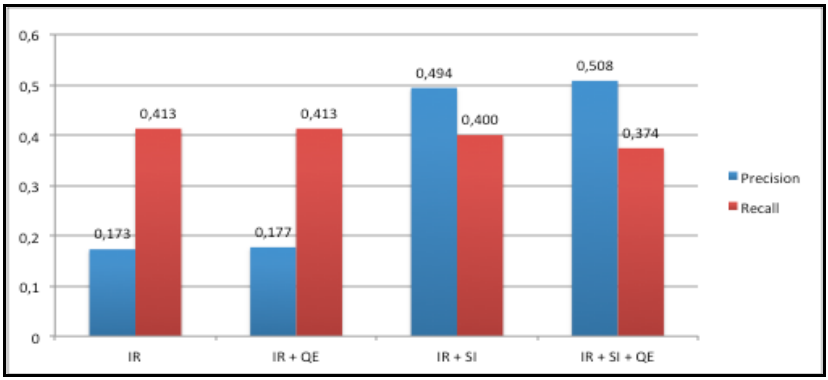


Figura 41: Resultados obtidos pelos diferentes métodos de RI.

Analisando a Figura 41, pode-se perceber que os métodos de expansão de pesquisa melhoram a precisão quando se comparam com as tradicionais técnicas de RI. O método tradicional retorna uma média de *precision* e *recall,* respectivamente de 0,173 e 0,413. Aplicando o método de expansão de pesquisa, as métricas de *precision* e recall, não mudaram o resultado. Isso se deu porque o usuário efetuou sua busca utilizando uma única palavra-chave em sua pesquisa. Ao aplicar a métrica que Casbral (2010)(Cabral, 2010) chamou de aprimoramento semântico, as métricas subiram para 0,494 de *precision* e 0,400 de *recall.* Esse aprimoramento semântico consiste em efetuar uma comparação entre a consulta do usuário com um dicionário fonético da língua portuguesa, a fim de encontrar registros relevantes na base de conhecimento. O conjunto resultante é comparado com a consulta gerada pelo método de expansão de pesquisa com a finalidade de retornar o maior número possível de resultados relevantes ao usuário.

No quarto teste, foram concatenadas as técnicas de IR com aprimoramento semântico e com a expansão de pesquisa. A união desses três métodos resultou em um *precision* de 0,508 e um *recall* de 0,374. Dessa forma, pode ser verificado que os resultados foram em parte, satisfatórios, visto que a precisão média aumentou de 0,177 para 0,508.

Já os índices de *recall* diminuíram de 0,413 para 0,374. Isso se deve ao fato de que a quantidade de documentos relevantes recuperados aumentou bastante.

7.2 SEGUNDO EXPERIMENTO – DETECÇÃO DE EXPRESSÕES NEGATIVAS

Para avaliação do sistema de detecção de expressões negativas foi executado primeiramente o algoritmo de detecção na base anotada pelo especialista, com o objetivo de avaliar a robustez do modelo proposto. O modelo exige que a partir da definição de um corpus anotado, sejam efetuados um conjunto de treinamento e um conjunto de testes na base em que se deseja recuperar a informação. Nesse caso, não foi efetuado o conjunto de testes, mas simplesmente foi aplicado o algoritmo na base anotada pelo especialista. A Tabela 4 apresenta as respostas do algoritmo.

Tabela 4: Resultado do sistema de detecção de expressões negativas em TC

Algoritmo / **Especialista**	**Afirmativo**	**Hipotético**	**Negativo**
Afirmativo	140	2	-
Hipotético	-	-	-
Negativo	-	-	19

Nesse conjunto de testes o sistema não considerou as sentenças que foram anotadas pelo especialista como “Ambíguo” e “Tema em destaque não é um achado”, por entender que nessas sentenças não são importantes para o sistema de detecção de frases realmente negativas. Esses casos serão analisados no trabalhos futuros para melhorar o sistema.

O algoritmo, nesse caso, retornou todas as sentenças que foram consideradas negativas pelo especialista, como sendo negativas. Entretanto, o algoritmo considerou duas sentenças como sendo hipotéticas e o especialista considerou essas sentenças como sendo afirmativas. As sentenças definidas pelo sistema como sendo hipotéticas foram essas:

- **PROVÁVEL** ANEURISMA DE "ACM" DIREITA.
- **PROVÁVEL** ARACNOIDOCELE SELAR ANTERIOR (VARIANTE)

O algoritmo detectou dois achados com precisão. A partir deles ele encontrou o termo “provável” e inferiu que a expressão deveria ser considerada hipotética. Entretanto, o especialista considerou essa expressão como sendo afirmativa. Após uma nova consulta ao especialista agora de posse do exame do paciente, esse especialista afirmou que a expressão é sim hipotética, pois ele não pode afirmar com precisão se o achado estava realmente presente no paciente. Nesses dois casos pode-se perceber que o especialista definiu uma expressão hipotética como sendo afirmativa e o sistema corrigiu o provável erro do especialista.

Como resultado, o sistema atingiu uma precisão de 0,9647 e um recall de 1,000 sem considerar que o especialista definiu erroneamente alguns achados. O mesmo procedimento de teste foi efetuado para os laudos de US. O resultado da detecção de expressões negativas em laudos de US pode ser visualizado na Tabela 5.

Tabela 5: Resultado do sistema de detecção de expressões negativas em US

Algoritmo / **Especialista**	**Afirmativo**	**Hipotético**	**Negativo**
Afirmativo	69	1	1
Hipotético	6	1	-
Negativo	0	-	53

O mesmo acontece para as sentenças de ultrassonografia, onde o sistema não considerou as sentenças que foram anotadas pelo especialista como “Ambíguo” e “Tema em destaque não é um achado”, por entender que nessas sentenças não são importantes para o sistema de detecção de frases realmente negativas. Esses casos serão analisados no em trabalhos futuros para melhorar o sistema.

O algoritmo retornou todas as sentenças que foram consideradas negativas pelo especialista, como sendo negativas. Entretanto, o algoritmo considerou duas sentenças como sendo hipotéticas e o especialista considerou essas sentenças como sendo afirmativas. As sentenças definidas pelo especialista como sendo hipotéticas foram essas:

1. VESÍCULA BILIAR MEDINDO DE CONTORNOS LISOS E REGULARES, CONTEÚDO HOMOGÊNEO, **SEM EVIDÊNCIA** DE LITÍASE.
2. OBSERVA-SE PEQUENA IMAGEM HIPERECOGÊNICA NO TERÇO MÉDIO DO RIM DIREITO.

3. ESTUDO ULTRASSONOGRÁFICO SUGESTIVO DE TIREOIDITE COM NÓDULOS.
4. OBSERVA-SE PEQUENA IMAGEM HIPERECOGÊNICA NO TERÇO MÉDIO DO RIM DIREITO, MEDINDO 3,8MM X 2,1MM, QUE **PODE CORRESPONDER** A CÁLCULO OU INTERFACE VASCULAR.
5. NEFRECTOMIA DIREITA.

A sentença 1 foi definida erroneamente pelo especialista como afirmativo, mas o sistema detectou como sendo uma expressão negativa. Na sentença 2, o sistema detectou erroneamente como sendo uma sentença hipotética, pois ele considerou o termo "pequena" como sendo hipotético. Para a expressão de número 3 o especialista considerou afirmativo e o sistema, hipotético. Analisando a expressão, o algoritmo considerou o termo "sugestivo" como sendo hipotético. Nesse caso, é extremamente difícil precisar se o termo disparador da expressão se refere ao estudo ou ao achado (nódulos).

Para as outras quatro expressões onde o especialista considerou como sendo hipotética, o sistema considerou como sendo afirmativa. Isso se deve ao fato que o especialista considerou a frase inteira e não somente a sentença. Por exemplo, na frase de número 4 o sistema reconheceu três sentenças, sendo que para as duas primeiras, foram consideradas afirmativas pelo sistema e a última, como hipotética. Nesse caso, o sistema corrigiu o erro do especialista, já que o estudo era reconhecer expressões negativas em cada sentença e não na frase/laudo como um todo (onde o especialista considerou o laudo por completo).

E para as outras frases, a exemplo da expressão de número 5, só havia duas palavras o sistema considerou como sendo afirmativas e o especialista considerou hipotética. Analisando essa expressão, não se pode entender o motivo pelo qual o especialista definiu como uma expressão hipotética.

Como resultado, o sistema atingiu uma precisão de 0,9642 e um recall de 0,9000 sem considerar que o especialista definiu erroneamente alguns achados.

7.3 ESTUDO DE CASO

Nessa seção, o objetivo é observar o comportamento do sistema em situações reais de pesquisa por um profissional de saúde. Nesse cenário, o objetivo é o usuário efetuar pesquisas por laudos que contenham determinados termos que são conhecidos na ontologia, por laudos que contenham expressões negativas e por laudos que contenham expressões que não são conhecidas pela ontologia. O objetivo desse estudo é mostrar ao usuário que o protótipo apresenta respostas ao usuário em suas consultas mesmo que o termo que ele estiver procurando não seja encontrado na base de laudos. Se mesmo assim o protótipo não encontrar os resultados que o usuário necessita, espera-se que o termo seja registrado para uma indexação futura. Essa informação é de extrema validade, pois o sistema poderá "aprender" novos termos conforme os usuários vão utilizando o modelo. E para os usuários, eles podem aprender sobre os termos mais usados por outros profissionais.

Para ilustrar as consultas executadas nesse estudo de caso, considere as seguintes *queries*:

Q1: "Presença de nódulos na tireóide"

Q2: "Ausência de litíases"

Q3: "Calcificação em pipoca no cérebro"

Q4: "Aneurisma de ACM frontal direita"

A Query Q1 foi efetuada com termos que estão presentes na ontologia de domínio. Nesse caso, os documentos esperados como resposta para a Q1 são todos os documentos que contenham "presença nódulos tireóide" + "Presença Nódulo Glândula Tireóide" + "presença Neoplasias Glândula Tireóide" + "presença Doenças Glândula Tireóide" + presença Glândula Tireóide" (resultado da expansão da pesquisa). Como o termo "nódulos na tireóide" é conhecido no DeCS, o sistema retornou todos os documentos que tinham alguma relação com o conjunto de termo listado na *query* expandida. Ainda, os resultados não podem conter expressões negativas, pois o usuário não selecionou o campo específico para procurar frases com sentido negativo. Um trecho do resultado em formato XML de Q1 pode ver visto na Figura 42.

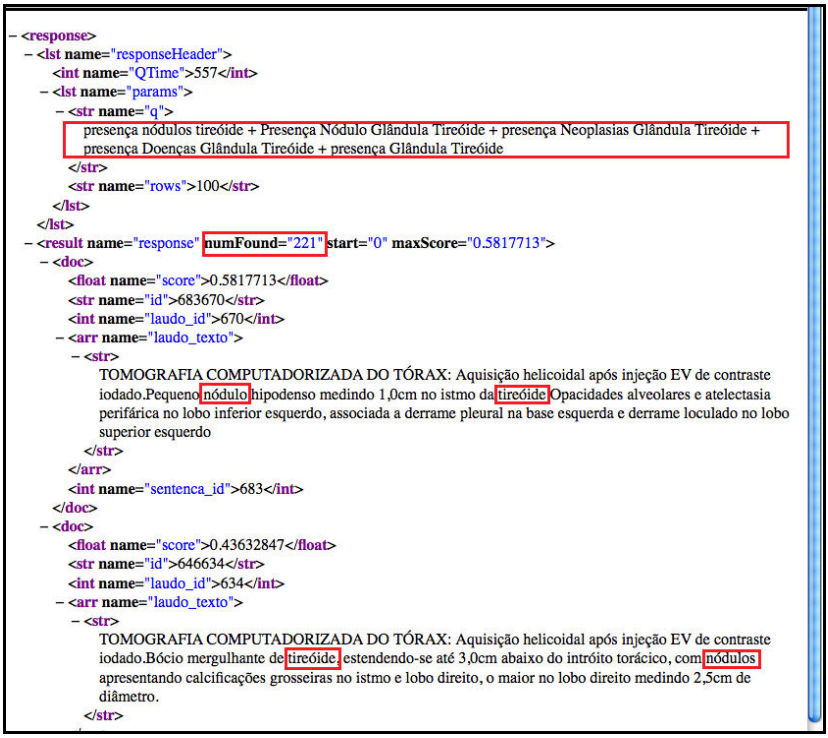


Figura 42: Exemplo de resultados para Q1.

A lista apresenta os laudos que contem somente expressões afirmativas e que tenham relevância com os termos solicitados. A pesquisa retornou 221 resultados, mas como no conjunto de laudos selecionados não continha nenhum laudo com a "presença de nódulos na tireóide", o protótipo trouxe os laudos que continham os termos em separado, mas que continham alguma similaridade. O resumo da pesquisa pode ser visualizado no Quadro 4, onde são apresentados a quantidade de laudos retornados, a precisão dos dez primeiros termos e o tempo de resposta para efetuar a consulta.

A pesquisa Q1 obteve em um primeiro momento um índice de acerto em P@10 de 50%. Após uma análise mais criteriosa nos resultados obtidos, pode-se perceber que vários laudos continham expressões com o seguinte texto:

"Glândula TIREÓIDE de forma, contornos e densidade normais"

Mesmo que o laudo contivesse nódulos em alguma outra frase, o resumo do laudo continha um texto que invalidava a pesquisa e

retornando assim, laudos que aparentemente não continham o achado que o usuário estava procurando. Para melhorar o índice de precisão, foi adicionada ao dicionário de expressões negativas uma inferência chamada "*not Achado*". Essa inferência diz que um laudo que contenha um determinado achado e venha seguido de uma expressão "normal", não contém esse achado, sendo assim, o laudo é considerado negativo. Na segunda iteração, os termos que continham "normal" dentro de uma sentença que continha um achado foram excluídos das respostas e a precisão subiu consideravelmente (chegou a 70%). Entretanto, foi encontrado um laudo que continha um termo hipotético: *"provavelmente corresponde a nódulo mergulhante de tireóide"*. Nesse caso, não foi considerado esse termo como válido para a pesquisa.

A *query* Q2 foi executada contra o mecanismo de busca por termos que contenham a expressão "Ausência de litíases". Nesse caso, o usuário necessita encontrar laudos que não possuam litíases. A *query* expandida conterá as seguintes expressões: "litíases" + "Calculose" + "Calculose do Ureter" + "Litíase do Ureter" + "Litíase Ureteral". Ainda, o algoritmo vai adicionar aos termos procurados as expressões com sentido negativo, como: "Ausencia de" + "sem sinais de" + "sem evidência de", etc.

Como resposta, o sistema retornou 432 resultados e a P@10 atingiu 100% de acerto. O Quadro 4 apresenta o índice de precisão e a quantidade de respostas pelo modelo aqui proposto. Além disso, a Q2 só retornou laudos que continham a presença de disparadores negativos, que podem ser visualizados na Figura 43.

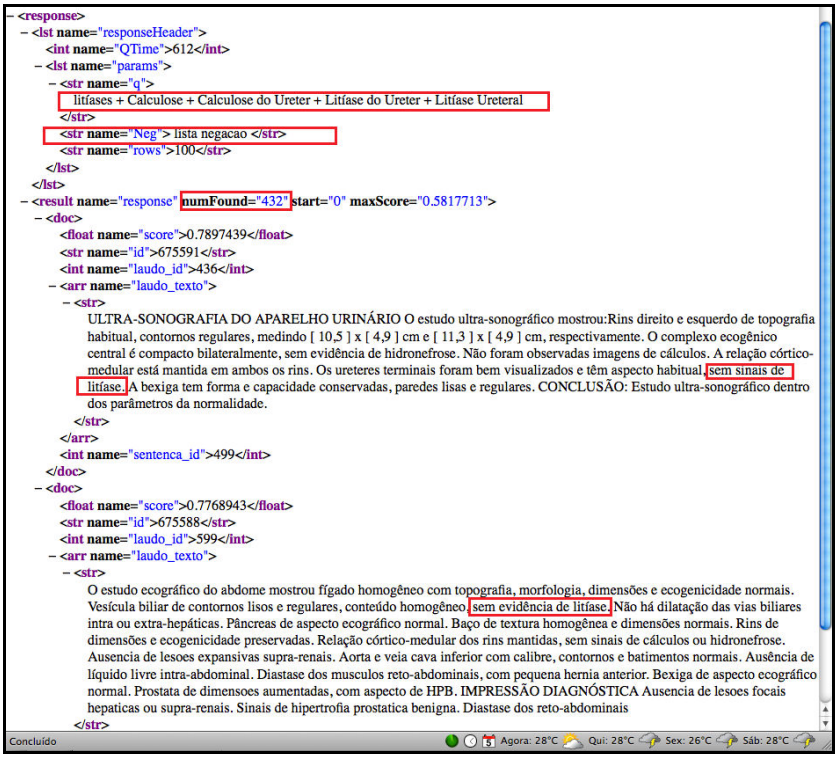


Figura 43: Exemplo de resultados para Q2.

Por outro lado, o sistema teve dificuldades em encontrar expressões negativas quando a explicitação do profissional médico ao prover seu lado, era dada em várias sentenças. Ou seja, quando o profissional utilizava várias frases para descrever seu raciocínio, muitas vezes o termo descrito na ontologia era divido em várias sentenças também. Por exemplo, ao analisar o seguinte laudo:

> O estudo US da **tireóide** mostrou pele e tecido celular subcutâneo preservados. **Tireóide** com topografia, morfologia, dimensões e textura normais. Não identificamos **nódulos**.

Nem o termo "tireóide", nem "nódulo" não constam na ontologia de forma isolada. A ontologia descreve o termo como sendo "Nódulo da Glândula Tireóide". Apesar do estudo acima estar 100% relacionado a esse termo da ontologia, o sistema não conseguiu anotar essa expressão e, dessa forma o achado "nódulos" não foi identificado pelo método de

anotação semântica e conseqüentemente, não foi recuperado pelo motor de busca, apesar de ter encontrado a expressão negativa.

Já as *queries* Q3 e Q4 são pesquisas que o usuário efetuou onde os termos são desconhecidos na ontologia. No caso de Q3, não foi possível identificar os documentos que continham as expressões solicitadas pelo usuário.

Para concluir a consulta ao usuário, o protótipo tentou efetuar a expansão da pesquisa, mas não encontrou nenhum termo na ontologia de domínio. Tentou encontrar termos sinônimos no dicionário da língua portuguesa, mas também não encontrou nada. Mesmo assim o sistema efetuou a busca com a expressão exata que o usuário solicitou, mas não encontrou nenhuma resposta. O termo foi então armazenado em uma base temporária e em um segundo momento o especialista verificou a relevância desse termo com o domínio da pesquisa e criou o relacionamento do termo com a ontologia.

O especialista referenciou o termo da consulta Q3, "calcificação em pipoca" com o descritor "neurocisticercose" na ontologia. O sistema registra esse conhecimento na base "termo_cotidiano" e a partir desse momento o sistema recupera os laudos que contem os termos "neurocisticercose" + "Cisticercose Encefálica" + "Cisticercose Cerebral" e "Cisticercose do Sistema Nervoso Central". Como resultado dessa ultima consulta, os laudos que contenham qualquer um desses termos são recuperados pelo mecanismo de busca. O Quadro 4 apresenta os resultados dessa consulta. A P@10 chegou a 90% e foram recuperados 109 laudos com a expressão expandida a partir da consulta inicial do usuário.

O mesmo acontece para a *query* Q4. Aqui o termo "ACM" é uma abreviação de "artéria cerebral média". A abreviação não consta na ontologia, mas a expressão por extenso, sim. Mesmo que o termo ACM não esteja referenciado à ontologia, esse termo é muito utilizado pela comunidade médica e as pesquisas por essa expressão retornaram 92 documentos. O termo ACM foi definido como sinônimo de artéria cerebral média para aumentar o índice de precisão do mecanismo de busca. Logo após o especialista ter referenciado o termo à expressão, o mecanismo de busca recupera os laudos que contenham os termos "ACM" e também "artéria cerebral média" em suas pesquisas. Como resultado, o sistema retornou 79 laudos que continham a *query* expandida juntamente com os critérios da Q4. A P@10, nesse caso, ficou em 100% .

Para validar esse estudo de caso e definir a precisão do sistema na recuperação das *queries*, foi utilizada a métrica de P@10. Para cada uma

das *queries* definidas no inicio dessa seção. O resultado é apresentado no Quadro 4. Para um melhor aproveitamento das pesquisas, as *queries* foram efetuadas após a validação e *linkagem* do especialista sobre os termos de uso cotidiano com a ontologia. A precisão média (*average precision*) de todas as *queries* do Quadro 4 foi de 0,9000.

Query	Resultado	Tempo em ms	p@10
Q1 "Presença de nódulos na tireóide"	221	557	0,7*
Q2: "Ausência de litíases"	432	612	1,0
Q3: "Calcificação em pipoca no cérebro"	109	736	0,9*
Q4: "Aneurisma de ACM frontal direita"	79	589	1,0
* Nessa *query* foram encontrados termos hipotéticos que não podem ser considerados como válidos			

Quadro 4: Consultas utilizando a metodologia desenvolvida.

Em seguida, foram executadas as mesmas *queries* no modelo tradicional de pesquisa para poder comparar com a tecnologia desenvolvida e aprimorar a precisão da recuperação conhecimento médico. O Quadro 5 apresenta as pesquisas efetuadas para poder chegar a uma resposta satisfatória pelo usuário.

Q1 "Presença de nódulos na tireóide"			
Query	**Resultado**	**Tempo em ms**	**p@10**
nódulo OR tireóide	592	531611	0,0
nódulo AND tireóide	17	296479	0,5
Presença AND nódulo AND tireóide	2	260205	0,2
Ausência AND nódulo AND tireoide	7	265321	0,0
Sem evidência AND nodulo AND tireóide	10	267713	0,2
Tireóide	52	295112	0,2
Tireóide AND NOT densidade normal	27	306273	0,1

Quadro 5: Consulta Q1 pelo método tradicional

Pode-se perceber que para conseguir um resultado satisfatório, o método tradicional exige um grande número de iterações e combinações de operadores booleanos. Por exemplo, quando pesquisado pela expressão "nódulo OR tireóide", os dez primeiros resultados não encontraram nenhum laudo que continham as respostas para a Q1. Em compensação para uma pesquisa que continha a expressão "nódulo AND tireóide", a P@10 teve um índice de acerto de 50%.

Para todas as outras pesquisas, o índice de precisão não passou de 0,2. Dessa forma, a precisão média de todas as respostas enviadas pelo sistema foi de 0,1714 em Q1. Mas para se obter todas as respostas e chegar à melhor *query,* foi necessário efetuar oito pesquisas na base de dados. Ou seja, o usuário perdeu 2.222.714 milissegundos, ou mais ou menos 37 minutos para conseguir uma resposta satisfatória do sistema, isso somente para a Q1.

A seguir foi executada a segunda *query*. Nessa pesquisa, o objetivo é encontrar laudos que não contenham o achado "litíase", para isso será necessário procurar por expressões negativas no laudo. O Quadro 6 apresenta as pesquisas efetuadas para poder chegar a uma resposta satisfatória pelo usuário na consulta Q2.

Q2 "Ausência de litíases"			
Query	**Resultado**	**Tempo em ms**	**p@10**
Ausência OR litíase	29284	3631500	0,0
Sem Evidência OR Litíase	724	611307	0,0
Ausência AND litíase	263	275378	0,2
Evidência AND litíase	188	250700	0,8
Sem evidência AND litíase	106	281820	0,8
Não há evidência AND litíase	53	265150	0,5

Quadro 6: Consulta Q2 pelo método tradicional.

A exemplo da consulta anterior, para essa nova pesquisa foram definidas seis *queries* até chegar a um resultado satisfatório. A primeira consulta obteve 29284 resultados e a segunda consulta obteve 724 resultados. Entretanto para as duas primeiras consultas, nenhum resultado atende aos critérios de consulta do usuário em virtude de ser uma pesquisa muito ampla e por não existir nenhum critério de

classificação das respostas. Mesmo que as respostas retornadas contenham o achado "litíase", os laudos que foram apresentados ao usuário, contem algum tipo de presença do achado e não a ausência dele.

Para a consulta "ausência AND litíase", o sistema encontrou somente 20% das respostas corretas em P@10. Esse fato se deu porque o médico especialista que proveu diagnóstico para exames com esse achado utilizava um modelo de laudo padrão que continha a expressão "ausência de lesões...". Mesmo a lesão não fazendo parte do termo "litíase, o buscador retornou alguns laudos que continham termos negados.

Já nas consultas "evidência AND litíase:" e "sem evidência AND litíase" o motor de busca não soube retornar os termos positivos ou negativos. Mesmo com uma diferença de 82 laudos para a consulta anterior, que não continham a expressão "sem evidência", o sistema retornou 80% dos laudos negativos. E para ultima consulta "não há evidência AND litíase", o motor de busca somente conseguiu precisar 50% dos laudos à Q2. Como resultado, a precisão média foi de 0,4166 em Q2 e consumiu do usuário um tempo de 5.315.855 milissegundos ou mais ou menos 88 minutos para conseguir chegar a uma resposta satisfatória.

Para a Q3, o objetivo é encontrar laudos que contenham os termos "Calcificação em pipoca no cérebro". O Quadro 7 apresenta as pesquisas efetuadas para poder chegar a uma resposta satisfatória pelo usuário.

Q3 - "Calcificação em pipoca no cérebro"			
Query	**Resultado**	**Tempo em ms**	**p@10**
Calcificação OR pipoca OR Cérebro	384	761703	0,0
Calcificação AND Cérebro	0	237531	0,0
Calcificação AND cranio	98	227926	0,3
Neurocisticercose	12	273051	0,6*
* Nessa *query* foram encontrados termos hipotéticos que não podem ser considerados como válidos			

Quadro 7: Consulta Q3 pelo método tradicional.

Para essa pesquisa, o mecanismo de busca teve a menor precisão de todas as outras consultas. Isso se deu ao fato de que o termo de uso

cotidiano não é utilizado pelos profissionais médicos quando eles emitem seus laudos. A pesquisa que teve o melhor índice de precisão foi como de era de se esperar a que tinha como palavra-chave o termo "neurocisticercose", com um índice de 60% de precisão. Entretanto, pode-se perceber que quando se procurava por laudos que continham expressões como "Calcificação AND crânio", o mecanismo de busca retornou 30% dos laudos como pertinentes à pesquisa do usuário. A precisão média desse conjunto de testes foi de 0,2250 em Q3, mas o usuário precisou de 25 minutos para conseguir encontrar a resposta corretamente.

A última consulta definida nesse estudo de caso foi a Q4 "Aneurisma de ACM frontal direita". O Quadro 8 apresenta os resultados encontrados quando efetuadas as pesquisas utilizando o método tradicional de busca utilizado no STT.

Q4 "Aneurisma de ACM frontal direita"			
Query	**Resultado**	**Tempo em ms**	**p@10**
Aneurisma OR ACM OR direita	38678	1771919	0,0
Aneurisma AND ACM AND Frontal AND direita	6	269888	0,5
Aneurisma AND ACM AND Direita	18	269676	0,4
Aneurisma AND ACM D.	10	282508	0,1
Aneurisma AND Artéria Cerebral Média AND direita	92	261257	0,5

Quadro 8: Consulta Q4 pelo método tradicional

Pode-se perceber que as consultas que obtiveram maiores resultados, obtiveram o melhor índice de precisão, excluindo a primeira consulta que procurou todos os termos utilizando o operador OR. Para a primeira consulta, como era de se esperar, retornou 38678 laudos e a precisão foi de 0%. Isso acontece porque as consultas com operador OR sempre retornam muitos resultados, são extremamente demoradas e os resultados quase sempre não são nada relevantes. No estudo de caso

aqui apresentado, todas as consultas com o operador OR resultaram em precisão de 0%.

As consultas que obtiveram melhores resultados em Q4 foram as que utilizaram todos os termos juntamente com o operador AND. A segunda consulta retornou somente seis termos e a precisão foi de 50%, ou seja, somente três dos seis laudos recuperados satisfaziam a solicitação do usuário. A melhor consulta foi a ultima, que utilizou o temo "artéria Cerebral média" ao invés da sua abreviação (ACM).

A precisão média de todas as pesquisas foi de 0,3000, mas para se chegar a um resultado satisfatório, o usuário precisou esperar 2.855.248, ou seja, quase 48 minutos para conseguir um resultado de 30% de precisão em suas pesquisas.

Para melhor visualização dos resultados obtidos foi efetuada uma comparação entre o tradicional método de pesquisa em IR contra a metodologia aqui proposta. A Figura 44 apresenta um gráfico com os resultados das consultas para as quatro *queries* descrita nesse estudo de caso.

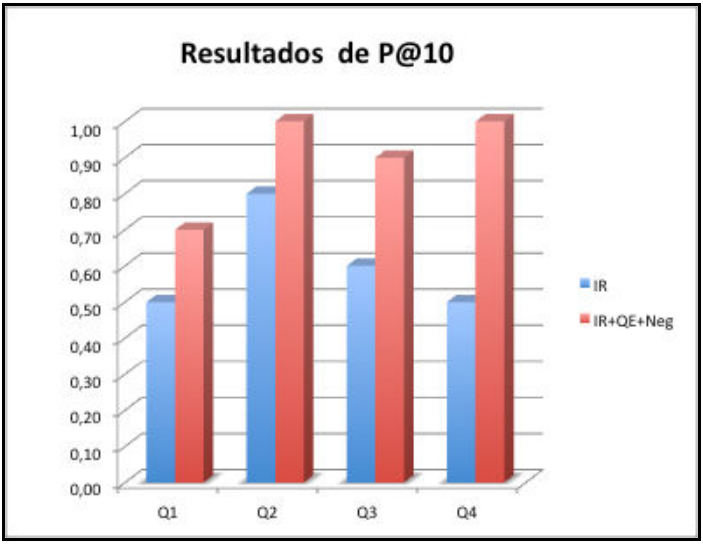


Figura 44: P@10 das consultas comparando IR com o modelo proposto.

A metodologia proposta avaliou a precisão dos dez primeiros resultados (P@10) para os dois modelos de pesquisa: o tradicional (aqui descrito como IR) e o novo modelo (IR+QE+Neg). Pode-se perceber que em todas as *queries* apresentadas, o novo modelo de pesquisa

obteve resultados muito melhores que o método tradicional. Para o usuário conseguir chegar a um resultado satisfatório usando mo método IR tradicional, ele necessitou efetuar diversas pesquisas. A Figura 44 comprara somente os melhores resultados do método tradicional com o IR+QE+Neg.

Com exceção da Q1, todas as outras consultas obtiveram uma precisão igual ou acima de 90%. O índice de precisão um pouco mais baixo em Q1 se deu pelo fato de nas respostas conterem muitas expressões hipotéticas. E dessa forma, não se pode confirmar a precisão dos laudos. Mas mesmo assim, a precisão da Q1 ainda foi muito superior ao método tradicional de IR (50%).

Da mesma forma, foi comparada a precisão média (*AveP*) das respostas dos dois modelos de pesquisa. Mesmo que o novo modelo de pesquisa retorne mais informações que o tradicional, a precisão é muito maior que o método tradicional. Os baixos índices de precisão média do método tradicional se deu em função da grande quantidade de informação que fora retornada. Conforme descrito pela fórmula de *AveP*, a precisão média utilizada nesse estudo de caso foi analisada em um universo de 50 termos, ou seja, somente as 50 primeiras respostas foram avaliadas para definir a *AveP*. A Figura 45 apresenta uma comparação entre o método tradicional de pesquisa e o modelo proposto.

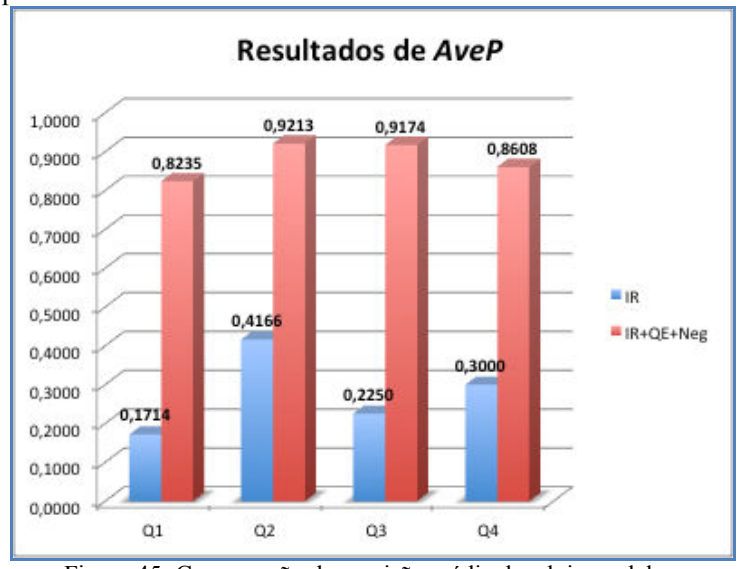


Figura 45: Comparação da precisão média dos dois modelos.

Como a precisão média dos dois modelos pode ser computada, pode-se ainda comparar esse novo modelo com os mais importantes apresentados na seção estado da arte. O Quadro 9 apresenta uma comparação entre o modelo proposto com os modelos analisados no estado da arte desta tese.

No. Sentenças	**352**	**50.000**	**1.000**	**1.058**	**558**	**352**
Modelos	IR Tradicional	Expansão de pesquisa		Detecção de Negação		Modelo Proposto
	Van Rijsbergen, 1975	**Díaz-Galiano et al., 2009**	**Abdou e Savoy, 2008**	**Chapman, 2001**	**Gindl et al., 2008**	**IR+ QE+Neg**
AveP	0,28	0,23	0,38	Não disponível	Não disponível	0,88
Precisão	0,30	Não disponível	Não disponível	0,78	0,68	0,96

Quadro 9: Comparação entre os modelos de pesquisa disponíveis na literatura.

Entretanto, a precisão média apresentada pelos autores, somente pode ser analisada nos métodos de expansão de pesquisa. Da mesma forma, a precisão total somente estava disponível nos trabalhos de detecção de expressões negativas. Como esse trabalho prevê a utilização de dois diferentes modelos, ficou difícil definir uma comparação eficiente contra os modelos disponíveis na literatura. Mesmo assim, o modelo aqui proposto apresentou uma precisão bem acima do que está disponível na literatura.

O modelo apresentado por Díaz-Galiano *et al.*, (2009)(Díaz-Galiano *et al.*, 2009) utiliza a base de dados do *Cross Language Evaluation Forun* (CLEF) do anos de 2005 e 2006. A base possui 50.000 imagens anotadas e que estão disponíveis para testes de precisão. Já os outros autores utilizaram uma base própria para medir a performance de seus esperimentos. A Figura 46 apresenta o comparativo entre os modelos pesquisados contra o modelo proposto.

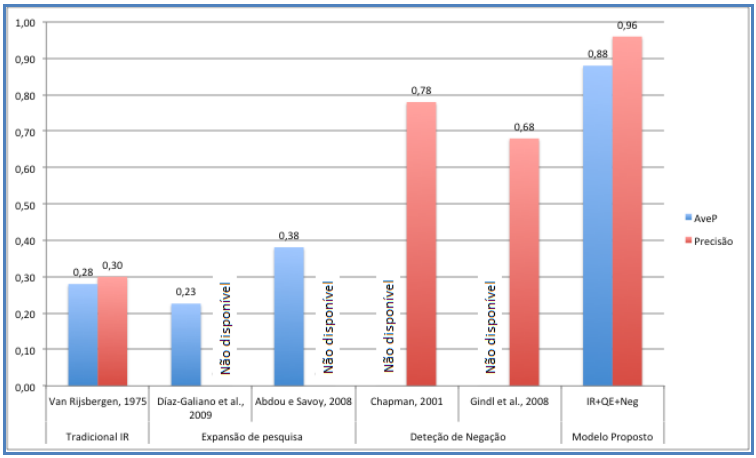


Figura 46: Comparação entre os modelos pesquisados.

Por fim, foram medidos os tempos de respostas dos dois modelos de pesquisa. O novo modelo aqui apresentado, utilizou a ferramenta do *lucene* para criar um índice invertido da base de conhecimento. Já o modelo tradicional de pesquisa não possui nenhuma forma de indexação dos dados. Por isso, a diferença de respostas foi extremamente grande. A Figura 47 apresenta um gráfico comparando os dois modelos para a obtenção das respostas das pesquisas. Os tempos são descritos em segundos.

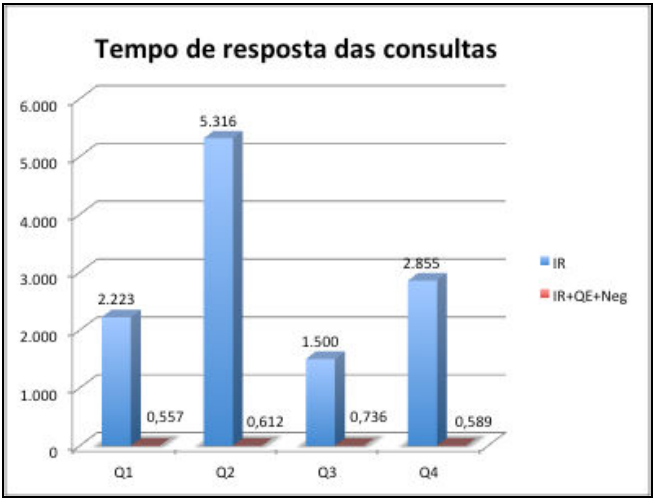


Figura 47: Comparação dos tempos de resposta dos dois modelos.

Esse gráfico descreve o tempo total em que o usuário necessitou para obter uma resposta satisfatória em suas consultas. Como o método tradicional exige do usuário várias iterações, o tempo para conseguir uma resposta foi extremamente alto. Mas mesmo que os dados estivessem indexados pelo banco de dados referencial, o usuário teve que repetir de quatro a oito vezes a sua pesquisa para obter as respostas do sistema. Já no novo modelo de pesquisa, o usuário efetuando somente uma única pesquisa já obtêm melhores resultados.

8 CONCLUSÕES E TRABALHOS FUTUROS

Nessa pesquisa foram discutidos os problemas relacionados à recuperação e comunicação do conhecimento a partir de bases de dados da área da saúde. Foram identificados problemas em que o usuário necessita lidar com (obter acesso a) uma grande quantidade de informações a fim de recuperar informações em registros médicos.

Destacou-se a dificuldade em encontrar determinada informação em bases de dados médicas e prover a comunicação do conhecimento adquirido no passado para utilizar em atendimentos futuros. Apesar de o domínio médico já possibilitar o acesso a ontologias e ferramentas de buscas para esse fim, essas ferramentas são de pouca utilidade quando utilizadas.

Baseado nesse problema de pesquisa foi definido como objetivo principal o desenvolvimento de um modelo que promovesse melhorias na indexação de um documento médico e considerável melhora no acesso a esses dados. Esse trabalho descreve um modelo que permite uma busca semântica a partir de ontologias médicas em um domínio específico, tais como textos em laudos médicos e toxicológicos. Métodos de arquitetura, indexação e consulta, foram discutidos durante o desenvolvimento dessa tese. Foi introduzido um modelo integrado de recuperação da informação (IRI), ou seja, uma abordagem conceitual nova para a RI, onde: informações de diagnóstico médico do paciente, informações do banco de dados toxicológicos baseados em ontologias e informações médicas de bibliotecas científicas, especialmente desenvolvidas para este trabalho, são recuperadas a partir desse modelo.

A revisão bibliográfica iniciou com a discussão do problema de acesso e representação do conhecimento, com foco em ontologias utilizadas na área da saúde. Diante disso, foi realizada uma extensa revisão na literatura em busca de informações sobre o assunto. Constatou-se que a área de recuperação da informação provê novas técnicas que poderiam ser utilizadas em conjunto para a melhoria do acesso aos documentos. A revisão da literatura mostrou que as técnicas de anotações semânticas e extração de expressões negativas, são utilizadas para extrair o conhecimento e proporcionar a RI. Foram identificados pontos fortes e fracos nos modelos de RI empregados na área médica. Cumpriu-se, pois, que a utilização de duas técnicas de RI, juntamente com o uso de anotações semânticas em ontologias, contribui para aperfeiçoar as pesquisas em documentos médicos, possibilitando a disseminação do conhecimento.

Durante o período de estudos de trabalhos relacionados aos objetivos dessa tese, não foi encontrada nenhuma pesquisa que abordasse a combinação das técnicas de anotações semânticas, expansão do conhecimento e detecção de expressões negativas para a indexação e representação do conhecimento em saúde, fato esse que originou a singularidade dessa pesquisa.

A mais importante contribuição dessa proposta de pesquisa é o desenvolvimento de um modelo para a indexação e representação do conhecimento na área da saúde. O uso dessa nova abordagem técnica possibilita maior acesso às tecnologias desenvolvidas, orientam os profissionais no atendimento de pacientes e permite a comunicação do conhecimento de forma a melhorar a transferência do conhecimento explicito entre os usuários do sistema.

O modelo de expansão do conhecimento utilizado nesse trabalho permite a extração de termos sinônimos da ontologia de domínio que não foram utilizados pelos profissionais durante sua pesquisa, para retornar mais informações a respeito do assunto pesquisado. Assim, o uso da ontologia tem se destacado na troca de informações entre sistemas computacionais, como no uso de, mas não somente, em sistemas de segunda opinião formativa e de suporte à decisão.

Também foi desenvolvida ontologia para a área de toxicologia clínica, a fim de classificar as informações disponíveis e estabelecer relações bem definidas entre essas informações (Cabral, 2010). A ontologia é um vocabulário controlado e um dicionário de sinônimos que possui uma diretriz de prática clínica para auxiliar o profissional médico nas decisões sobre cuidados de saúde adequados para o domínio toxicológico. O uso dessa ontologia aperfeiçoa o processo de atendimento, aumentando a confiabilidade e eficiência do profissional médico.

O uso da ontologia DeCS permite que não sejam encontradas nos laudos, informações que contenham sentido ambíguo. Pois o usuário é direcionado a utilizar os termos que estão na ontologia para representar o conhecimento explicito e gerar assim a base de informações de uso cotidiano. Esse conhecimento explícito do profissional é manifestado durante o uso do mecanismo de busca. Mas essa integração somente poderá ocorrer após a validação do termo pesquisado com o termo constante na ontologia, por um profissional capacitado.

A atualização da base de conhecimento nesse modelo é facilitada, pois, a identificação de novos termos é feita baseado na freqüência com que eles são pesquisados. Esse conhecimento auxilia, pois, na identificação de novas linguagens utilizadas pela comunidade médica

que ainda não foram incorporadas e disseminadas aos outros profissionais de saúde.

Para lidar com declarações negativas, foi desenvolvido um algoritmo de mineração de texto a partir da base de dados de diagnóstico médico do STT com a finalidade de extrair termos médicos. Esses termos foram utilizados para classificar o texto, como por exemplo, "o paciente não tem a hipertensão", "o paciente tem hipertensão", ou "o paciente tem pressão alta".

O uso de pesquisas que contenham expressões com sentido negativo permite que os usuários encontrem somente informações que dizem respeito àquela busca em questão. Por isso, o conceito aqui utilizado, proporciona ao usuário escolher pelo uso de incluir expressões negativas ou não em uma pesquisa. Como resultado, as informações mais precisas são enviadas aos utilizadores do sistema.

Entretanto, esse modelo possui algumas dificuldades em anotar informações da ontologia quando determinado assunto está dividido em duas ou mais frases. Nesse caso há a necessidade de se estudar técnicas de anotações semânticas baseadas no contexto.

O modelo de pesquisa desenvolvido permite que um usuário ou médico visualize as informações do paciente mais rápido que a tecnologia atual e os resultados mais significativos com o uso de busca semântica. Além disso, um profissional pode realizar buscas em laudos de exames de pacientes sobre um diagnóstico de um atendimento feito no passado que possam ajudar no tratamento futuro de um paciente, melhorando assim a qualidade no atendimento ao paciente.

8.1 SUGESTÕES PARA TRABALHOS FUTUROS

Com o objetivo de possibilitar uma melhor utilização desse modelo, pretende-se como primeiro trabalho futuro, fazer uma análise mais aprofundada junto a especialistas de domínio, a ampliação dessa tecnologia dentro do sistema de Telemedicina e Telessaúde. Dessa forma, mais usuários poderão utilizar o sistema e contribuir para o aperfeiçoamento do modelo. Pode-se ainda prover o uso de expressões regionalistas, para determinar quais termos são mais utilizados em pesquisas e assim criar novos cursos de capacitação profissional baseado nas dúvidas dos usuários.

Há a necessidade de efetuar experimentos mais extensos a fim de medir os benefícios da recuperação e representação do conhecimento

apresentado nesse modelo de tese. O objetivo aqui será analisar os resultados recuperados pela ferramenta de busca em busca de resultados falsos positivos e/ou falsos negativos. Permitir avaliar possíveis benefícios no contexto da utilização de expressões de uso cotidiano no dia-a-dia do profissional médico também é um dos objetivos desse novo estudo.

Outra consideração, diz respeito ao mecanismo de anotação semântica. O método de anotação possui alguns erros para referenciar o termo à ontologia. Por exemplo, o termo "pós" foi anotado pela ontologia DeCS como produto químico, mas no laudo esse termo é definido como um prefixo significando "Depois de":

"Presença de moderado resíduo pós-miccional".

Esse erro pode ser facilmente removido se a fase de pré-processamento do texto considerar o esses termos como sendo uma *stopword*. Outro exemplo desse problema é que o analisador removeu o prefixo "pré" da pesquisa. E assim, termos como "pré-natal" não foram anotados pelo sistema. Esses casos devem ser tratados como especiais e o sistema não deve filtrá-los. Por isso há a necessidade de um estudo mais aprofundado no anotador semântico a fim de remover essas imprecisões.

Como informado anteriormente, o modelo de detecção de expressões negativas tem dificuldade em encontrar termos em sentenças diferentes, mas que estão relacionados. Dessa forma, há a necessidade de estudar outras formas de análise de texto baseado em contexto para poder anotar semanticamente a informação e permitir uma melhor detecção dos dados. Uma possível forma de resolver esse problema é utilizar a técnica de descoberta de conhecimento por associação entre características do texto à procura de padrões de informação (Feldman *et al.*, 1998).

REFERÊNCIAS

ABDOU, S.; SAVOY, J. Searching in Medline: Query expansion and manual indexing evaluation. *Inf. Process. Manage.*, v. 44, n. 2, p. 781-789, 2008.

AGOSTI, M.; BONFIGLIO-DOSIO, G.; FERRO, N. A historical and contemporary study on annotations to derive key features for systems design. *International Journal on Digital Libraries*, v. 8, n. 1, p. 1-19, 2007.

ALAG, S. *Collective intelligence in action*. Manning, 2008.

ANDRADE, R.; COMUNELLO, E.; RIBEIRO, L. A.; WANGENHEIM, A. V. An Approach to Semantic Search im Medical Databases. In: Proceedings of 8th International Information and Telecommunication Technologies Symposium, Florianópolis. Fundação Bardal de Educação e Cultura, 2009.

ANSI/NISOZ39-19-2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies. . n. 01 de dezembro de 2010. Bethesda: NISO Press: American National Standards Institute, 2005.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern information retrieval*. Addison-Wesley Reading, MA, 1999.

BASILI, V.; CALDIERA, G.; ROMBACH, D. The goal question metric approach. In: MARCINIAK, J. (Ed.). *Encyclopedia of Software Engineering*: Wiley, 1994.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, v. 284, n. 5, p. 34-43, 2001.

BHOGAL, J.; MACFARLANE, A.; SMITH, P. A review of ontology based query expansion. *Inf. Process. Manage.*, v. 43, n. 4, p. 866-886, 2007.

BODENREIDER, O.; BURGUN, A. *Aligning knowledge sources in the UMLS: methods, quantitative results, and applications*. 2004. (Pt 1).

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, v. 30, n. 1-7, p. 107-117, 1998.

BUCKLEY, C.; VOORHEES, E. M. Evaluating evaluation measure stability. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*. Athens, Greece: ACM, 2000. p. 33-40.

CABRAL, R. B. *Concepção, implementação e validação de um enfoque para integração e recuperação de conhecimento distribuído em bases de dados heterogêneas.* (2010). Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianopolis, 2010.

CHAPMAN, W. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, v. 34, n. 5, p. 301-310, 2001.

CHAPMAN, W.; BRIDEWELL, W.; HANBURY, P.; COOPER, G.; BUCHANAN, B.; CHAPMAN PHD, W.; BS, W.; BS, P.; COOPER MD PHD, G.; BUCHANAN PHD, B. Evaluation of Negation Phrases in Narrative Clinical Reports. 2002. Disponível

em:<*http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.5296*>.

CLUNIE, D. *DICOM structured reporting*. PixelMed Publishing, 2000.

CLUNIE, D. A. DICOM Standart. 2008. Disponível em:<*http://www.dclunie.com/dicom-status/status.html*>. Acesso em: 15 de janeiro de 2008.

DAQING, H.; DAN, W. Enhancing query translation with relevance feedback in translingual information retrieval. *Information Processing & Management*, v. 47, n. 1, p. 1-17, 2011.

DECS. Descritores em Ciências da Saúde. *BIREME/OPAS*, 2010. Disponível em:<*http://decs.bvs.br/I/decswebi2009.htm* >. Acesso em: 15 de maio de 2010.

DÍAZ-GALIANO, M. C.; MARTÍN-VALDIVIA, M. T.; UREÑA-LÓPEZ, L. A. Query expansion with a medical ontology to improve a multimodal information retrieval system. *Computers in Biology and Medicine*, v. 39, n. 4, p. 396-403, 2009.

DUMAS, M.; ALDRED, L.; HERAVIZADEH, M.; TER HOFSTEDE, A. Ontology markup for web forms generation. *Relation*, v. 10, n. 1.105, p. 3331, 2007.

EGC. Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento. n. 03 de Janeiro de 2011, 2011. Disponível em:<*http://www.egc.ufsc.br/index.php?option=com_content&view=article&id=7&Itemid=13&lang=pt*

>. Acesso em: 03 de Janeiro de 2011.

EMBLEY, D.; DING, Y.; LIDDLE, S.; VICKERS, M. Automatic creation and simplified querying of semantic Web content: An approach based on information-extraction ontologies. In: In Proceedings of the first Asian Semantic Web Conference (ASWC 2006) LNCS 4185. Citeseer, 2006. p.400-414.

FALOUTSOS, C.; OARD, D. A survey of information retrieval and filtering methods. *University of Maryland at College Park, College Park, MD*, 1995.

FELDMAN, R.; DAGAN, I.; HIRSH, H. Mining Text Using Keyword Distributions. *Journal of Intelligent Information Systems*, v. 10, n. 3, p. 281-300, 1998.

FILETO, R. Pesquisa Semântica., n. 01/02/2011, 2011. Disponível em:<*http://www.inf.ufsc.br/~fileto*>. Acesso em: 01/02/2011.

GILLELAND, M. Levenshtein distance, in three flavors. *Merriam Park Software*, 2009. Disponível em:< *http://www.merriampark.com/ld.htm*>. Acesso em: 07 de setembro de 2009.

GINDL, S.; KAISER, K.; MIKSCH, S. Syntactical negation detection in clinical practice guidelines. *Studies in health technology and informatics*, v. 136, p. 187, 2008.

GREENGRASS, E. Information retrieval: A survey. *University of Maryland, Baltimore County*, 2001. Disponível em:<*http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.1855*>. Acesso em: 20 de Novembro de 2009.

GROOTJEN, F. A.; VAN DER WEIDE, T. P. Conceptual query expansion. *Data & Knowledge Engineering*, v. 56, n. 2, p. 174-193, 2006.

GSCHWANDTNER, T.; KAISER, K.; MARTINI, P.; MIKSCH, S. Easing semantically enriched information retrieval--An interactive semi-automatic annotation system for medical documents. *International Journal of Human-Computer Studies*, v. 68, n. 6, p. 370-385, 2010.

HATCHER, E.; GOSPODNETIC, O. *Lucene in Action (In Action series)*. Manning Publications Co., 2004.

HRISTIDIS, V.; FARFÁN, F.; BURKE, R.; ROSSI, A.; WHITE, J. Information Discovery on Electronic Medical Records. Citeseer, 2007.

HUANG, H. Enterprise PACS and image distribution. *Computerized Medical Imaging and Graphics*, v. 27, n. 2-3, p. 241-253, 2003.

JONES, K.; WALKER, S.; ROBERTSON, S. *A probabilistic model of information retrieval: development and status*. Technical Report TR-446, Cambridge University Computer Laboratory, 1998.

KIRYAKOV, A.; POPOV, B.; OGNYANOFF, D.; MANOV, D.; KIRILOV, A.; GORANOV, M. Semantic Annotation, Indexing, and Retrieval. *The SemanticWeb - ISWC 2003*, 2003a. p. 484-499.

_____. Semantic Annotation, Indexing, and Retrieval.*The SemanticWeb - ISWC 2003*, 2003b. p. 484-499.

KONG, G. L.; XU, D. L.; YANG, J. B. Clinical Decision Support Systems: A review on knowledge representation and inference under uncertainties. *International Journal of Computational Intelligence Systems*, v. 1, n. 2, p. 159-167, 2008.

KOTSAKIS, E. Structured information retrieval in XML documents. *Proceedings of the 2002 ACM symposium on Applied computing*. Madrid, Spain: ACM, 2002. p. 663-667.

LANDRY, R.; AMARA, N.; PABLOS-MENDES, A.; SHADEMANI, R.; GOLD, I. The Knowledge-Value Chain: A Conceptual Framework for Knowledge Translation in Health. *Bull World Health Organ*, v. 84, n. 8, p. 597-602, 2006.

LASSILA, O.; MCGUINNESS, D. The role of frame-based representation on the semantic web. *Electronic Transactions on Artificial Intelligence*, v. 6, n. 5., 2001. Disponível em:<*http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.125.4297*>.

LIN, K. H.-Y.; HOU, W.-J.; CHEN, H.-H. Retrieval of Biomedical Documents by Prioritizing Key Phrases In: Proceedings of the 14th Text REtrieval Conference, Gaithersburg, Maryland. 2005.

LOURENÇO, A.; CARREIRA, R.; GLEZ-PEÑA, D.; MÉNDEZ, J. R.; CARNEIRO, S.; ROCHA, L. M.; DÍAZ, F.; FERREIRA, E. C.; ROCHA, I.; FDEZ-RIVEROLA, F.; ROCHA, M. BioDR: Semantic indexing networks for biomedical document retrieval. *Expert Systems with Applications*, v. 37, n. 4, p. 3444-3453, 2010.

MAIA, R. S.; WANGENHEIM, A. V.; NOBRE, L. F. A Statewide Telemedicine Network for Public Health in Brazil. *Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems*: IEEE Computer Society, 2006. p. 495-500.

MANICA, H. *Modelo de Recuperação e Comunicação de Conhecimento em Emergência Médica com Utilização de Dispositivos Portáteis*. (2009). 155 f. (PHD Thesis) - Programa de Pós Graduação

em Engenharia e Gestão do Conhecimento, UFSC, Florianópolis, 2009. Disponível em:< *http://btd.egc.ufsc.br/?p=292*>.

MANICA, H.; ROCHA, C. C. D.; TODESCO, J. L.; DANTAS, M. A. R.; BAUER, M. A. Toward Developing Knowledge Representation in Emergency Medical Assistance through a Ontology-based Semantic Cache Model. *Proceedings of the The 21st International Conference on Software Engineering and Knowledge Engineering,*. Boston, USA, 2009. p. 592-596.

MANNING, C. D.; RAGHAVAN, P.; SCHTZE, H. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

MARON, M. E.; KUHNS, J. L. On Relevance, Probabilistic Indexing and Information Retrieval. *J. ACM*, v. 7, n. 3, p. 216-244, 1960.

MCNEILL, K.; WEINSTEIN, R.; HOLCOMB, M. Arizona telemedicine program. *Journal of the American Medical Informatics Association*, v. 5, n. 5, p. 441, 1998.

MOSKOVITCH, R.; SHAHAR, Y. Vaidurya: A multiple-ontology, concept-based, context-sensitive clinical-guideline search engine. *Journal of Biomedical Informatics*, v. 42, n. 1, p. 11-21, 2009.

MUNIR, K.; ODEH, M.; MCCLATCHEY, R.; KHAN, S.; HABIB, I. Semantic Information Retrieval from Distributed Heterogeneous Data Sources. *FIT Islamabad, special track on bioinformatics for academia and industry*, Dec. 21-22, 2006.

MUTALIK, P. G.; DESHPANDE, A.; NADKARNI, P. M. Use of general-purpose negation detection to augment concept indexing of medical documents: a quantitative study using the UMLS. *Journal of the*

American Medical Informatics Association : JAMIA, v. 8, n. 6, p. 598-609, 2001.

MYKOWIECKA, A.; MARCINIAK, M.; KUPSC, A. Rule-based information extraction from patients' clinical data. *J. of Biomedical Informatics*, v. 42, n. 5, p. 923-936, 2009.

ORENGO, V. M.; HUYCK, C. Relevance feedback and cross-language information retrieval. *Inf. Process. Manage.*, v. 42, n. 5, p. 1203-1217, 2006.

REEVE, L.; HAN, H. Survey of semantic annotation platforms. *Proceedings of the 2005 ACM symposium on Applied computing*. Santa Fe, New Mexico: ACM, 2005. p. 1634-1638.

ROKACH, L.; ROMANO, R.; MAIMON, O. Negation recognition in medical narrative reports. *Information Retrieval*, v. 11, n. 6, p. 499-538, 2008.

ROSARIO, B. Latent semantic indexing: An overview. [Techn. rep.]. 2000. Disponível em:<*http://www.sims.berkeley.edu/rosario/projects/LSI.pdf*>. Acesso em: 20 de janeiro de 2011.

RUBIN, D. L.; SHAH, N. H.; NOY, N. F. Biomedical ontologies: a functional perspective. *Briefings in Bioinformatics*, v. 9, n. 1, p. 75-90, January 1, 2008 2008.

SAGER, N.; FRIEDMAN, C.; LYMAN, M. S. *Medical Language Processing: Computer Management of Narrative Data*. Addison-Wesley Longman Publishing Co., Inc., 1987.

SAUVAGNAT, K.; BOUGHANEM, M.; CHRISMENT, C. Answering content and structure-based queries on XML documents using relevance propagation. *Information Systems*, v. 31, n. 7, p. 621-635, 2006.

SAVOY, J. An extended vector-processing scheme for searching information in hypertext systems. *Information Processing & Management*, v. 32, n. 2, p. 155-170, 1996.

SILVERSTEIN, C.; MARAIS, H.; HENZINGER, M.; MORICZ, M. Analysis of a very large web search engine query log. *SIGIR Forum*, v. 33, n. 1, p. 6-12, 1999.

SINGHAL, A. Modern information retrieval: A brief overview. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering.*, v. 24, n. 4, p. 35-43, 2001.

SNOMED. Systematized Nomenclature of Medicine. *SNOMED International. A division of the College of American Pathologists (CAP).* 2010. Acesso em: 07 de janeiro de 2010.

SPINK, A.; WOLFRAM, D.; JANSEN, M.; SARACEVIC, T. Searching the web: The public and their queries. *Journal of the American Society for Information Science and Technology*, v. 52, n. 3, p. 226-234, 2001.

STROHMAN, T. *Efficient processing of complex features for information retrieval.* (2007). Ph.D. Dissetation, University of Massachusetts Amherst, 2007.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, v. 25, n. 1-2, p.161-197, 1998. Disponível em:<*http://www.sciencedirect.com/science/article/B6TYX-*

3SYXJ6S-G/2/67ea511f5600d90a74999a9fef47ac98>. Acesso em: 01 ago 2005.

TIE. Telemedicine Information Exchange. 2009. Disponível em:<*http://tie.telemed.org/articles/article.asp?path=telemed101&article=tmcoming_nb_tie96.xml*>. Acesso em: 24 de agosto de 2009.

TURTLE, H.; CROFT, W. B. Inference networks for document retrieval. *Proceedings of the 13th annual international ACM SIGIR conference on Research and development in information retrieval*. Brussels, Belgium: ACM, 1990. p. 1-24.

VAN RIJSBERGEN, C. J. *Information retrieval / C. J. van Rijsbergen*. London ; Boston :: Butterworths, 1975. (Accessed from *http://nla.gov.au/nla.cat-vn1157566)*.

WALLAUER, J.; MACEDO, D.; ANDRADE, R.; VON WANGENHEIM, A. Building a National Telemedicine Network. *IT Professional*, v. 10, n. 2, p. 12-17, March/April 2008.

WANGENHEIM, A.; JUNIOR, C.; WAGNER, H.; CAVALCANTE, C. Caminhos para a Implantação de Telemedicina em Larga Escala: A Experiência de Santa Catarina. *Latin American Journal of Telehealth, América do Norte*, 3, abr. 2010. 2010.

WILLIAM, B. F.; RICARDO, B.-Y. (Eds.) **Information retrieval: data structures and algorithms**: Prentice-Hall, Inc., p.504ed. 1992.

APENDICE A - Publicações

Título: An Approach to Semantic Search in Medical Databases.
Evento: 8th International Information and Telecommunication Technologies Symposium, 2009
Local: Florianópolis –SC
Autores: ANDRADE, R.; COMUNELLO, E; RIBEIRO, L. A.; WANGENHEIM, A.v.
Estrato Qualis: B4

Título: Semantic Information Indexing and Retrieval on Patient Medical Data.
Evento: 8th International Information and Telecommunication Technologies Symposium, 2009
Local: Florianópolis.
Autores: ANDRADE, R.; CABRAL, R. B.; BARCELLOS JUNIOR, C. L.; WANGENHEIM, A.v
Estrato Qualis: B4

Título: Asyncronous Data Replication: A National Integration Strategy for Databases on Telemedicine
Evento: The 21th IEEE International Symposium on Computer-Based Medical Systems, 2008.
Local: Jyväskylä - Finlândia
Autores: ANDRADE, R.; MACEDO, D. D. J.; WANGENHEIM, Aldo Von; PERANTUNES, H. W. G.; DANTAS, M. A. R
Estrato Qualis: A2

Título: Busca Semântica Aplicada a Informações Clínicas
Evento: Congresso Brasileiro de Informática em Saúde, 2008
Local: Campos do Jordão - SP.
Autores: BARCELLOS JUNIOR, C. L.; ANDRADE, R.; RIBEIRO, L. A.; WANGENHEIM,
Estrato Qualis: B5

Título: Uma Ontologia Para o Atendimento Emergencial de Pacientes.
Evento: XI Congresso Brasileiro de Informática em Saúde (CBIS'08), 2008.

Local: , Campos do Jordão - SP.
Autores: LOPES, P. M. A; ANDRADE, R.; WANGENHEIM, A.v
Estrato Qualis: B5

Título: Plataforma de Gerência do Conhecimento Aplicada em um Ambiente de Toxicologia Clínica e Toxicovigilância
Evento: XI Congresso Brasileiro de Informática em Saúde (CBIS'08), 2008.
Local: Campos do Jordão – SP.
Autores: CABRAL, R. B.; SAVARIS, A.; ANDRADE, R.; ZANNIN, M.; WANGENHEIM, A.v..
Estrato Qualis: B5

APÊNDICE B – Sistema Catarinense de Telemedicina e Telessaúde

O Portal de Telemedicina possui 4 perfis de acesso, além do acesso do paciente:

1. **Técnico**

Os técnicos podem enviar os exames de duas formas distintas:

a. Manual: o exame é realizado em um equipamento médico local e "exportado" para formato digital. O técnico envia para o portal de telemedicina o exame do paciente. Este exame é protocolado digitalmente por um servidor de Protolação Digital de Documentos Eletrônicos (PDDE) e somente após o exame é armazenado no banco de dados;

b. Automatizado: um equipamento de aquisição de imagens que possui exportação em formato digital DICOM, como um tomógrafo computadorizado, envia os exames diretamente para um servidor localizado na instituição que gerou o exame. Esse servidor locar é chamado de Bridge DICOM. Esta Bridge, que é responsável pelo armazenamento temporário das imagens, realiza o envio das imagens recebida diretamente dos aparelhos para o servidor DICOM Central. O servidor DICOM, ao receber o exame de um determinado paciente, converte automaticamente as imagens para o formato JPG e armazena no banco de dados central os dois formatos de imagens (JPG para visualização rápida na web e o original DICOM).

2. **Requisitante**

Um médico com perfil de requisitante pode visualizar os exames e imagens de seus pacientes. Este médico pode também imprimir essas informações, gerar um protocolo para acesso do paciente.

3. **Executor**

Um médico com o perfil de "executor" pode acessar os exames de pacientes publicados no portal e efetuar um laudo para cada exame publicado. Cada médico executor possui também uma especialidade e somente poderá visualizar e emitir um laudo dos exames de sua especialidade. Este médico também pode emitir uma segunda opinião de um exame que foi laudado por outro profissional médico da mesma especialidade. O médico também pode conferir um laudo efetuado por outro profissional médico com o mesmo perfil de acesso e especialidade.

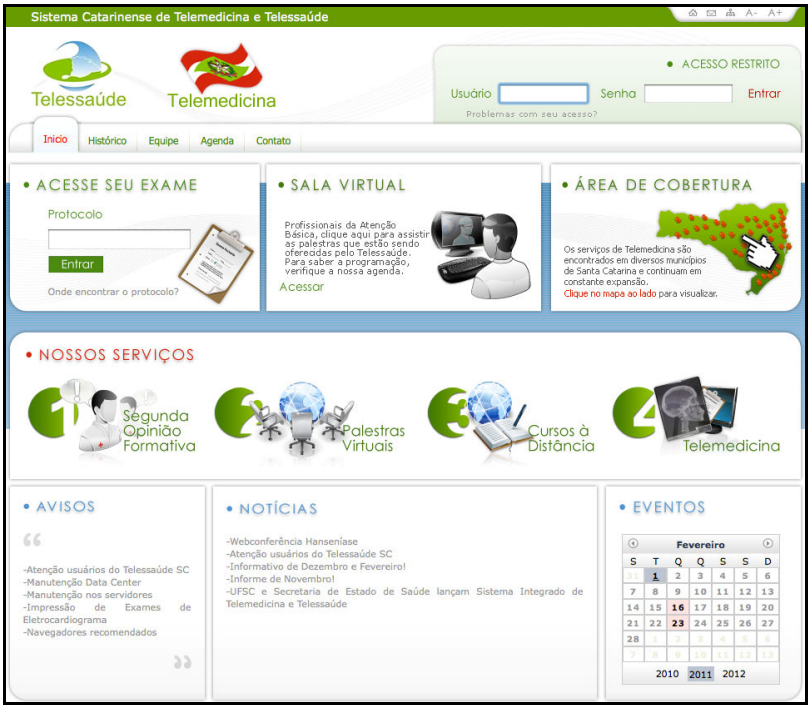
4. **Regulador**

Um médico com o perfil de Regulador tem acesso total aos exames enviados ao portal. Esse perfil é utilizado somente pelos médicos da Secretaria de Saúde de Santa Catarina (SES/SC) para poder efetuar o controle de atendimento de todos os hospitais e clínicas conveniadas com o Sistema Único de Saúde Pública brasileiro.

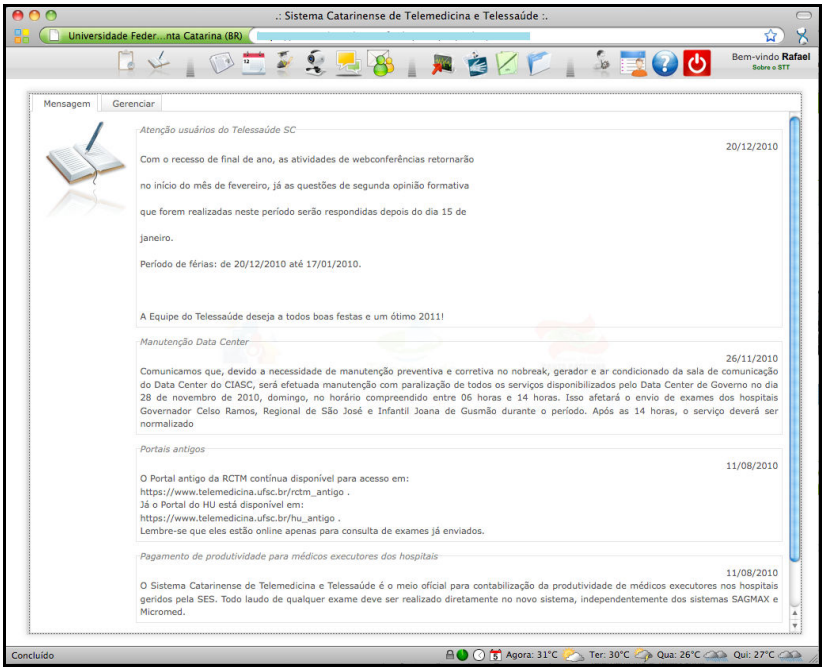
Por exemplo, quando um paciente necessitar de uma intervenção cirúrgica, os médicos da SES/SC devem antes de autorizar o procedimento, verificar se realmente o paciente necessita de tal procedimento operatório. Neste caso, os médicos com perfil "regulador" necessitam visualizar todos os exames efetuados pelo paciente para poder autorizar ou não o procedimento. E em alguns casos, esses médicos podem solicitar uma segunda opinião em um exame com resultado duvidoso, ou solicitar que o paciente efetue outros exames para que eles possam ter certeza que o paciente realmente necessita de uma intervenção cirúrgica.

Após a execução do laudo, o médico que requisitou o exame ao paciente poderá acessar digitalmente o(s) resultado(s) deste(s) exame(s) diretamente na web. O paciente também recebe um aviso informando que seu(s) exame(s) encontra(m)-se disponível(is) para acesso. E com uma senha e um protocolo de acesso o paciente poderá acessar seu exame diretamente na Web.

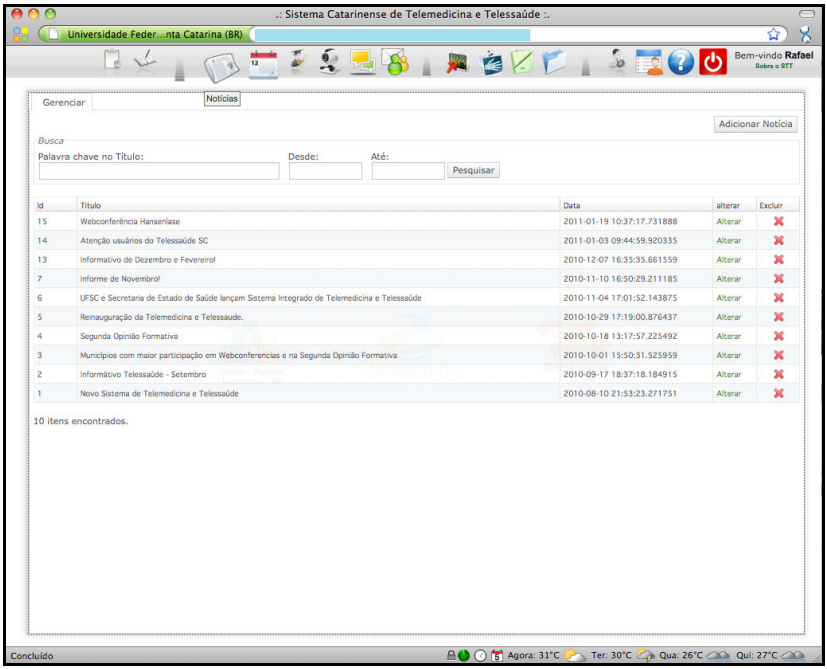
Abaixo são apresentadas as telas principais do STT:



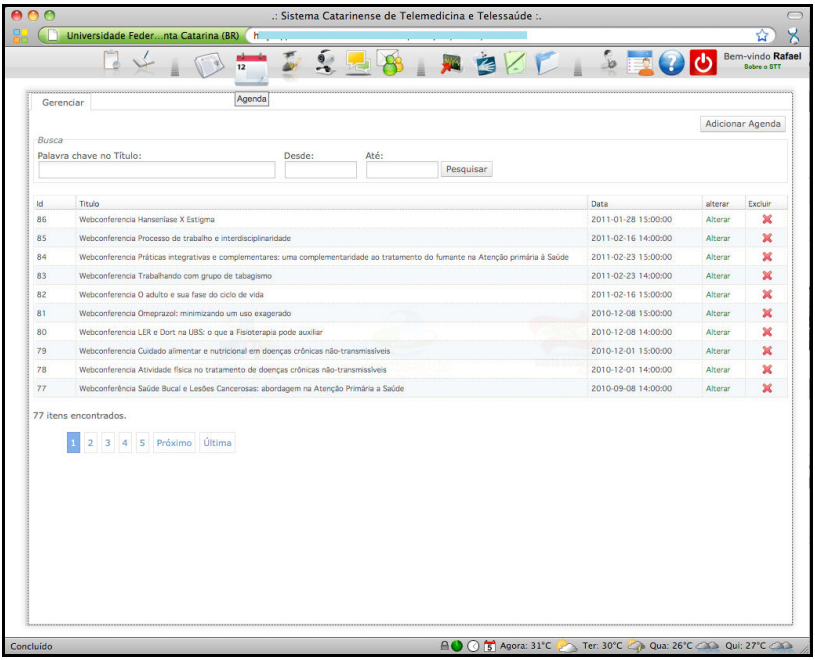
Tela Inicial do sistema



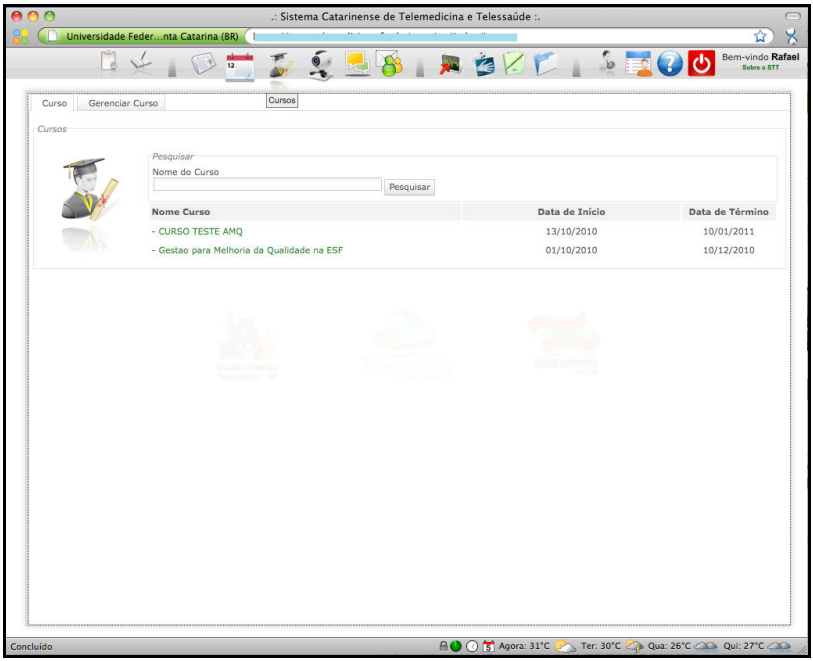
Tela de mensagens do sistema.



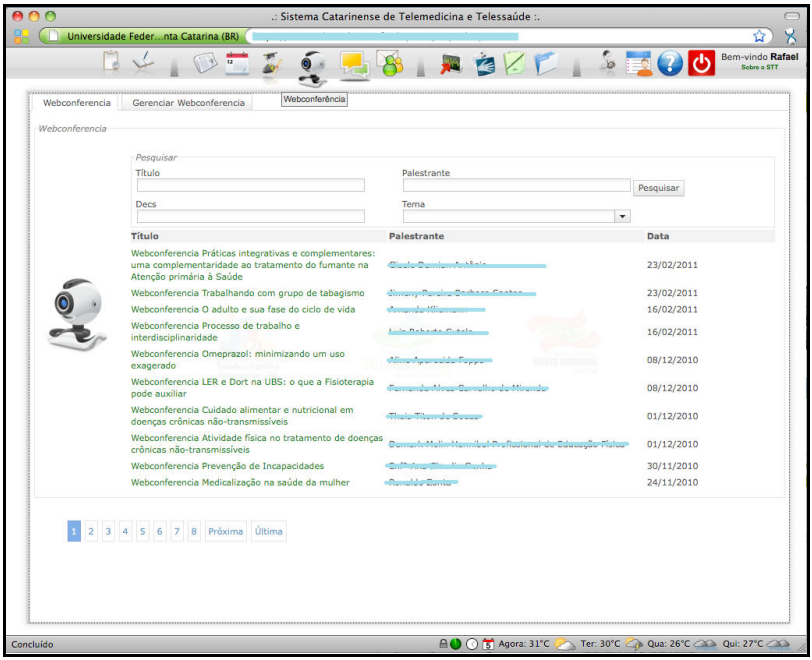
Tela de Notícias.



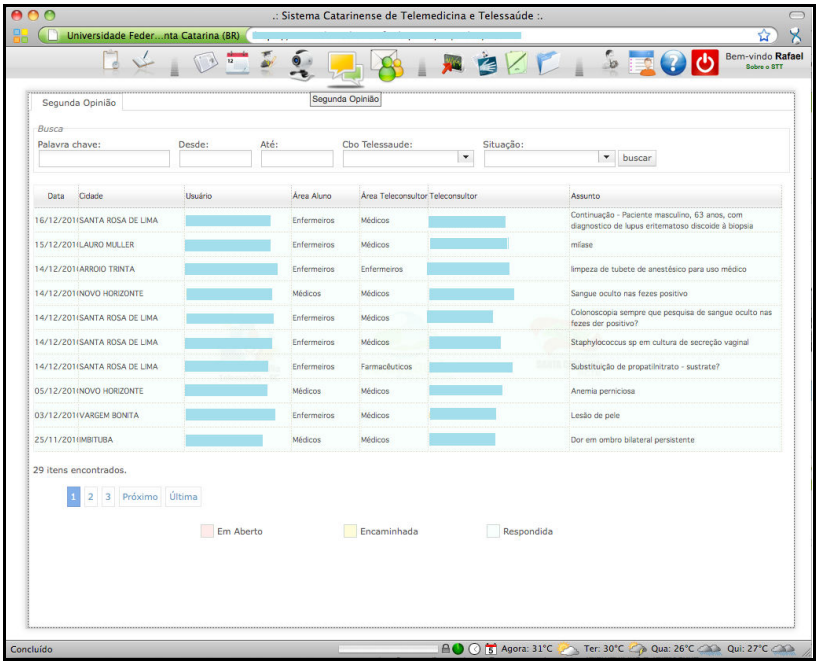
Tela de agenda do usuário conectado ao sistema



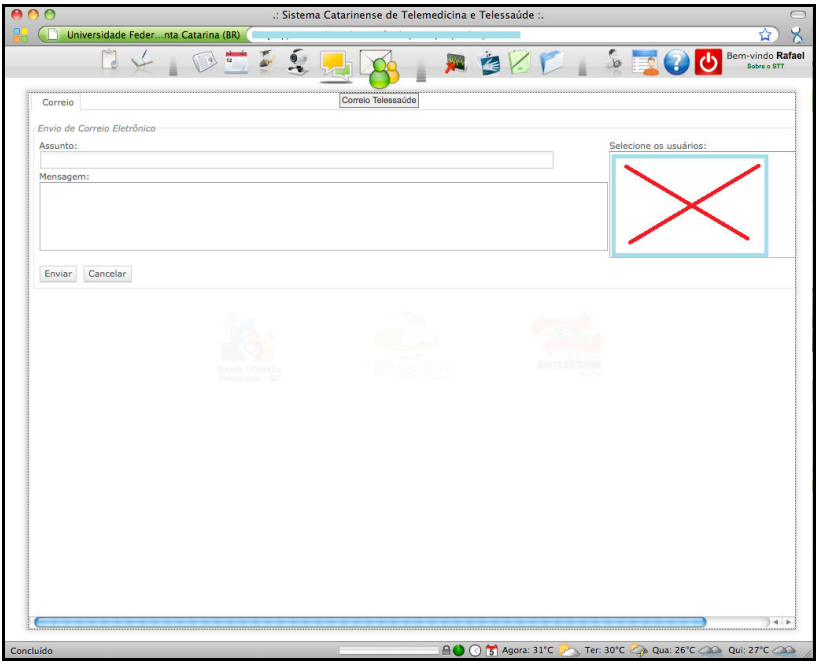
Cursos disponíveis ao usuário



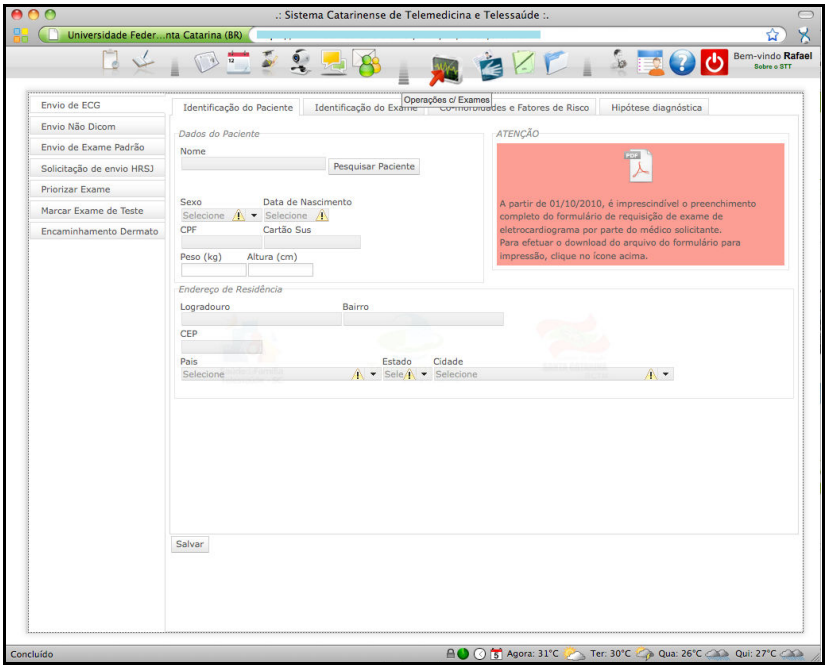
Tela de Webconferências.



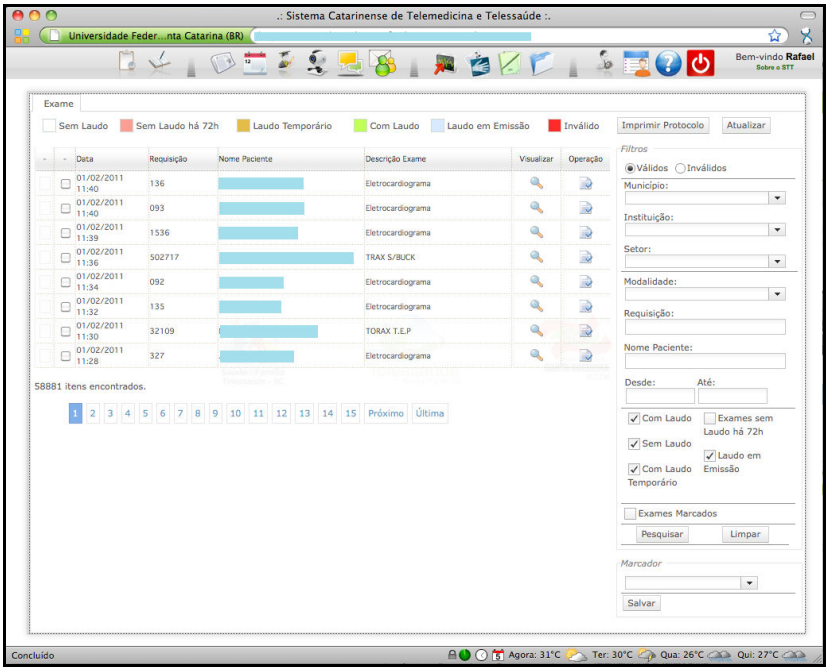
Tela de segunda opinião formativa.



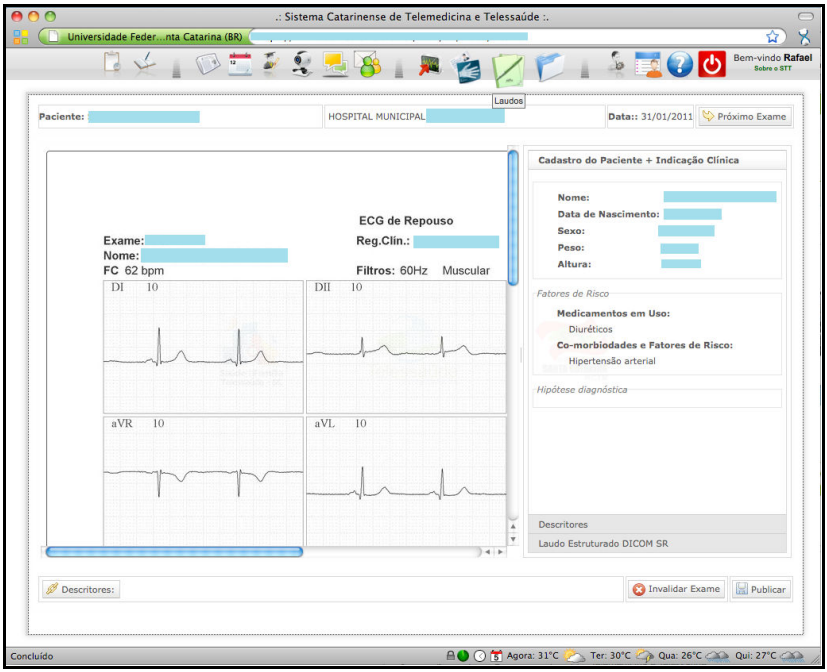
Tela de correio do telessaúde. Utilizado para troca de mensagens entre os usuários do sistema.



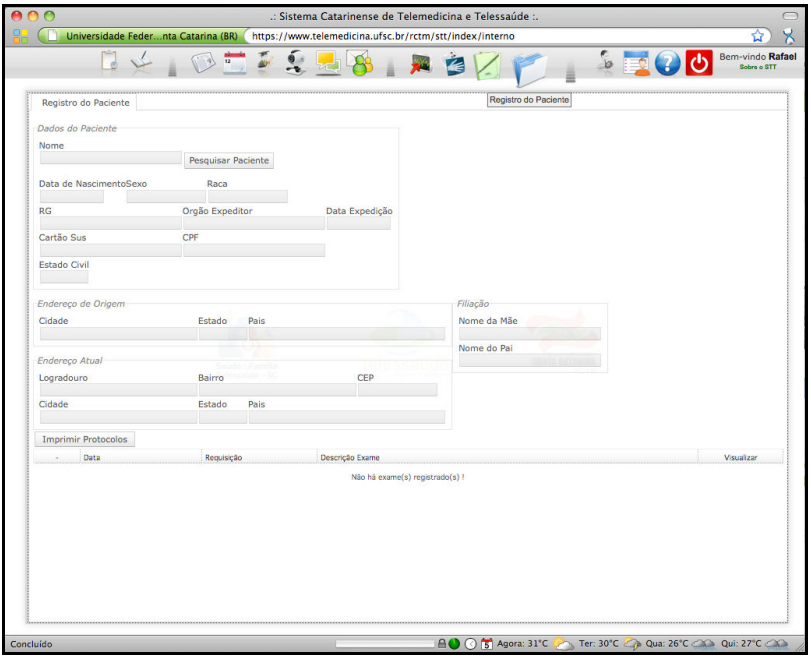
Tela de exames. Utilizada para cadrastro de paciente e envio de exames para o Sistema central.



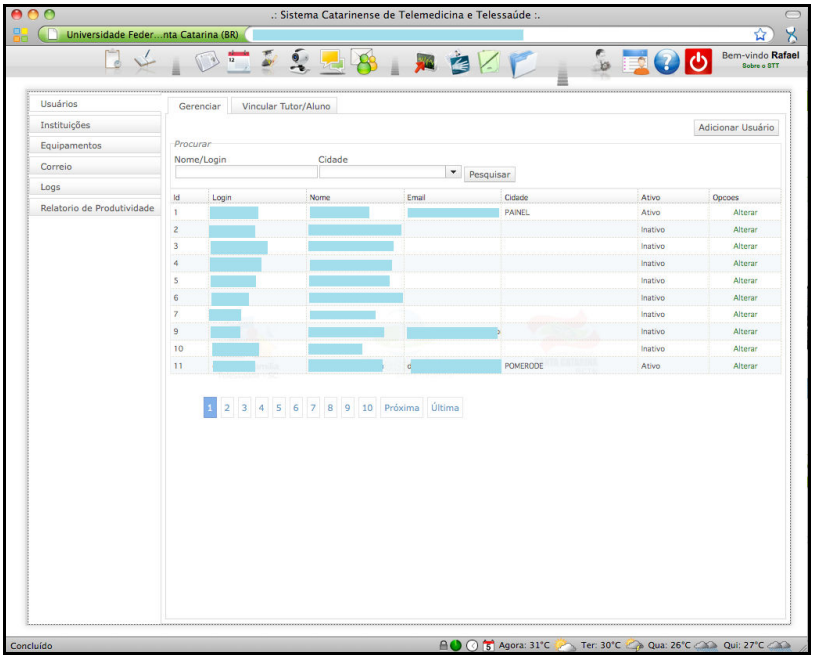
Tela de visualização de exames e filtros para exames.



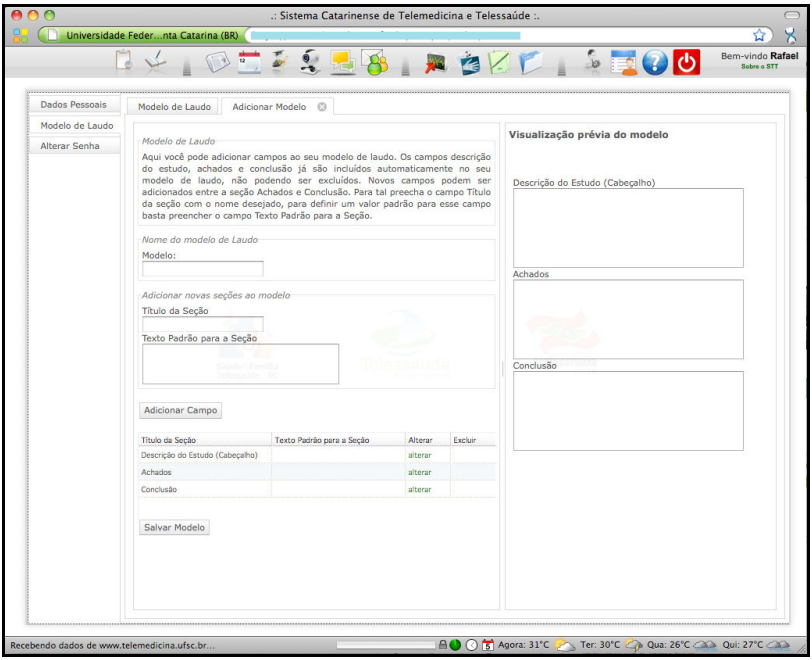
Tela de laudos



Tela de registro de pacientes.



Tela administrativa do sistema.



Tela de dados pessoais do usuário. Nessa tela um usuário médico pode cadastrar seus modelos de laudos e visualizar seus dados pessoais. Para os outros usuários só estão disponíveis os dados pessoais.