UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ARTHUR SCARDINI DOMINGUES

MOTOR DE BUSCA PARA INTEGRAÇÃO DE DADOS SOBRE SAÚDE NA WEB

MOSSORÓ/RN

2019

ARTHUR SCARDINI DOMINGUES

MOTOR DE BUSCA PARA INTEGRAÇÃO DE DADOS SOBRE SAÚDE NA WEB

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação – associação ampla entre a Universidade Federal Rural do Semi-Árido e a Universidade do Estado do Rio Grande do Norte, para a obtenção do título de Mestre em Ciência da Computação.

Orientadora: Angélica Félix de Castro, D.Sc – UFERSA

Coorientador: Francisco Milton Mendes Neto, D.Sc – UFERSA

MOSSORÓ/RN

2019

D671m DOMINGUES, ARTHUR.
MOTOR DE BUSCA PARA INTEGRAÇÃO DE DADOS SOBRE SAÚDE NA WEB / ARTHUR DOMINGUES. - 2019.
96 f. : il.

Orientadora: ANGELICA FELIX.
Coorientador: FRANCISCO MILTON MENDES.
Dissertação (Mestrado) - Universidade Federal Rural do Semi-árido, Programa de Pós-graduação em Ciência da Computação, 2019.

1. Saúde. 2. Motor de Busca. 3. Integração. 4. Base de Dados. 5. Ontologia . I. FELIX, ANGELICA, orient. II. MENDES, FRANCISCO MILTON, co-orient. III. Título.

ARTHUR SCARDINI DOMINGUES

MOTOR DE BUSCA PARA INTEGRAÇÃO DE DADOS SOBRE SAÚDE NA WEB

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação para a obtenção do título de Mestre em Ciência da Computação.

APROVADA EM: 27 / 03 / 2019.

Angélica Félix de Castro
Profa. Dra. Angélica Félix de Castro
Orientadora e Presidente da Banca

Prof. Dr. Francisco Milton Mendes Neto
Co-Orientador e Membro Interno - UFERSA

Aquiles M.F. Burlamaqui
Prof. Dr. Aquiles Medeiros Filgueira Burlamaqui
Examinador Externo – UFRN

Profa. Dra. Cicília Raquel Maia Leite
Examinadora Interna – UERN

Dedico este trabalho a quem mais se esforçou na sua realização: **EU**, sem demagogia nem Hipocrisia . . .

AGRADECIMENTOS

Primeiramente a Deus, pela vida e por me fortalecer em momentos em que eu mesmo já não encontrava saída. Neste, que é o maior desafio da minha vida, até hoje.

A meus pais Dilcemari e André Luiz, por todo amor e carinho, por me criarem e sempre me darem tudo aquilo que eu nunca soube dar valor.

A minha namorada Gizely, por ser minha companheira em todos momentos. Por me mostrar um lado da vida, que nunca imaginei conhecer.

Ao amigo Robson Locatelli, o desbravador do desconhecido. Pelo incentivo e conselhos até chegar a terra prometida.

Aos amigos da UFERSA e UERN. Companheiros na luta diária que passaram pelas mesmas dificuldades. Vocês se tornaram exemplos de luta, todo sucesso a nós.

A todos que conviveram comigo no condomínio Residencial Genésio Xavier 2. A vida em comunidade me ensinou que viver sozinho é pequeno demais. Um agradecimento em especial a Lucas, Edgar, Igor, Jesaias, Kaynara, Heloísa, Ítalo, Ramon, Mateus e Letícia. Muito Obrigado por Tudo!

A minha filha Victoria. Por você, um sentimento que eu nunca senti por ninguém. Cresça minha flor, o mundo é pequeno pra você.

A todos amigos que fiz durante minha caminhada em terras potiguares. Muito obrigado por me acolherem.

A meu co-orientador, Francisco Milton Mendes Neto, por todo esclarecimento e direcionamento em minha pesquisa.

A minha orientadora Angélica Félix de Castro, por todos ensinamentos e puxões de orelha.

"... Deus te deu a inteligência, não as respostas para suas perguntas".

Danilo Teixeira

RESUMO

A busca por informações sobre saúde vem se tornando uma atividade cada vez mais recorrente no cotidiano da população. Mediante essa busca, é possível para o indivíduo entender sobre o seu atual estado de saúde e, consequentemente, obter uma melhoria na qualidade de vida. Atualmente, o meio de comunicação mais utilizado para busca de informações é a internet. A internet possibilita a ampliação desse entendimento devido à grande quantidade de informações nela disponível, porém, em contrapartida, em meio a essa ampla quantidade de conteúdo, existem informações inadequadas ou incompletas. Desta forma, a utilização de motores de busca pode ajudar a fornecer conteúdos de qualidade, visto que são ferramentas desenvolvidas para vasculharem informações de acordo com critérios estabelecidos pelos usuários. Para isso, é necessário que o motor esteja inserido em um domínio onde as fontes que disponibilizam essas informações sejam repositórios confiáveis, ou seja, com alto índice de credibilidade. O presente trabalho apresenta a criação de um motor de busca sobre dados de saúde na web. O objetivo do motor é disponibilizar informações de bases de dados médicas, para aplicações computacionais de saúde. O motor faz uso de uma ontologia para classificação de conteúdo. A ontologia funciona como um filtro para a grande quantidade de dados recuperados pelo motor, a fim de fornecer somente informações relevantes ao contexto da busca.

PALAVRAS-CHAVE: Saúde, Motor de Busca, Integração, Base de Dados, Ontologia

ABSTRACT

The search for health information has become an increasingly recurrent activity in the daily lives of the population. Through this search, it is possible for the individual to understand about their current state of health and consequently to obtain an improvement in quality of life. Currently, the most used media information search is the internet. The Internet makes possible the extension of this knowledge due to the large amount of information available, but in contrast to this large amount of content, there is inadequate or incomplete information. Search engines can help provide quality content because they are tools designed to seek information according to criteria set by users. For this, it is necessary that the engine should inserted in a domain where the sources that make the information available are reliable repositories, that is, with a high index of credibility. This paper proposes the creation of a search service on health data on the web. The present work presents the creation of a search service on health data on the web. The purpose of the service is to provide information from medical databases for health computing applications. The service makes use of an ontology for content classification. The ontology functions as a filter for the large amount of data retrieved by the engine in order to provide only relevant information to the search context.

KEYWORDS: Health, Search Engine, Integration, Database, Ontology

LISTA DE ABREVIATURAS E SIGLAS

ADA	American Diabetes Association
API	Application Programming Interface
BVS	Biblioteca Virtual de Saúde
CI	Ciência da Informação
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
MESH	Medical Subject Headings
MIX	Metadata for Digital Still Images Standards Committee
MODS	Metadata Object Description Schema
NIH	National Institutes of Health
OWL	Ontology Web Language
PHP	Hypertext Preprocessor
PLN	Processamento de Linguagem Natural
PLUS	Picture Licensing Universal System
RDF	Resource Description Framework
RI	Recuperação da Informação
SGML	Standard Generalized Markup Language
SRI	Sistema de Recuperação da Informação
SWRL	Semantic Web Rules Language
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	eXtended Markup Language
XMP	Extensible Metadata Platform

LISTA DE FIGURAS

Figura 1 - Declaração de Tipo de Documento 24
Figura 2 - Declaração de Tipo de Instância 24
Figura 3 - Exemplo de Código HTML 25
Figura 4 - Exemplo de Código XML 27
Figura 5 - Exemplo de Tripla 28
Figura 6 - Sujeito e Predicado 29
Figura 7 - Representação em Código XML 30
Figura 8 - Tipos de Documentos 37
Figura 9 - Componentes Motor de Busca 39
Figura 10 - Fluxograma de Rastreamento 40
Figura 11 - Arquitetura do Motor 49
Figura 12 - Tela Inicial 50
Figura 13 - Tela de Cadastro 51
Figura 14 - Tela Principal 51
Figura 15 - Tela de Cadastro 52
Figura 16 – Site da BVS 53
Figura 17 – Arquivo XML Capturado 54
Figura 18 – Principais Classes da Ontologia 54
Figura 19 – Hierarquia de Classes 59
Figura 20 – Descrição da Classe Medication 59
Figura 21 – Classe DiabetsContent 61
Figura 22 – Classe HighRecommended..... 62
Figura 23 – Classe MediumRecommended..... 62
Figura 24 – Classe LowRecommended 62
Figura 25 – Telas MobiLEHealth 63
Figura 26 – Conexão com o Motor 64
Figura 27 – Tela de Conteúdo Cientifico 64
Figura 28 – Conteúdo Recomendado 65
Figura 29 – Classes Afirmativas 68
Figura 30 – Primeira Simulação 69
Figura 31 – Resultado Primeira Simulação 69
Figura 32 – Segunda Simulação 69

Figura 33 – Resultado Segunda Simulação 70
Figura 34 – Terceira Simulação 70
Figura 35 – Resultado Terceira Simulação 71
Figura 36 – Quarta Simulação 71
Figura 37 – Resultado Quarta Simulação 72

LISTA DE TABELAS

Tabela 1 - Elementos e Especificações Padrão Dublin Core 32
Tabela 2 - Exemplo de Documento 42
Tabela 3 - Índice Invertido 42
Tabela 4 - Análise de Competidores 48
Tabela 5 - Classificação de Conteúdos 49
Tabela 6 - Conteúdos Recuperados e Classificados 74

LISTA DE GRÁFICOS

Gráfico 1 - Faixa Etária dos Pesquisadores 74
Gráfico 2 - Leitura dos Conteúdos 74
Gráfico 3 - Termos Referentes ao Diabetes 75
Gráfico 4 - Assuntos Interessantes Abordados 75
Gráfico 5 - Nível de Recomendação 76
Gráfico 6 - Auxílio no Tratamento 76
Gráfico 7 - Clareza Quanto ao Tema 77
Gráfico 8 - Conteúdos Adequados 78
Gráfico 9 - Satisfação Quanto à Qualidade 78
Gráfico 10 - Melhor Qualidade de Vida 78
Gráfico 11 - Indicação de Conteúdos 79

SUMÁRIO

1. INTRODUÇÃO 17
1.1. CONTEXTO DE PESQUISA 17
1.2. MOTIVAÇÃO 18
1.3. OBJETIVOS 19
1.3.1. OBJETIVOS ESPECÍFICOS 19
1.4. METODOLOGIA 19
1.5. ESTRUTURA DO DOCUMENTO 21
2. REVISÃO BIBLIOGRÁFICA 22
2.1. FONTES DE DADOS DE SAÚDE 22
2.1.1. BASES DE DADOS DE SAÚDE 22
2.2. METADADOS 23
2.2.1. SGML 24
2.2.2. HTML 26
2.2.3. XML 27
2.2.4. RDF 29
2.2.5. PADRÕES DE METADADOS 32
2.2.6. PADRÃO DUBLIN CORE 33
2.3. RECUPERAÇÃO DA INFORMAÇÃO 34
2.3.1. MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO 35
2.3.2. AVALIAÇÃO DA RECUPERAÇÃO DA INFORMAÇÃO 38
2.4. MOTOR DE BUSCA 40
2.4.1. RASTREADOR 41
2.4.2. INDEXADOR 43
2.4.3. BUSCADOR 44
2.5. ONTOLOGIAS 44
2.5.1. METODOLOGIAS DE DESENVOLVIMENTO 46
2.6. TRABALHOS CORRELATOS 47
3. MOTOR DE INTEGRAÇÃO DE DADOS 50
3.1. ARQUITETURA DO MOTOR 50
3.2. INTERFACE DO MOTOR 51
3.3. BUSCA DE CONTEÚDOS 54

3.4. INDEXAÇÃO DE CONTEÚDOS 56
3.4.1. ONTOLOGIA PARA CLASSIFICAÇÃO 57
3.5. INTEGRAÇÃO 64
4. EXPERIMENTOS E VALIDAÇÕES 66
4.2. TESTES EXPERIMENTAIS 67
4.3. AVALIAÇÃO DA ONTOLOGIA 68
4.3.1. VERIFICAÇÃO 68
4.3.2. VALIDAÇÃO 73
4.4. QUESTIONÁRIO DE AVALIAÇÃO DE CONTEÚDOS 74
5. CONSIDERAÇÕES FINAIS 80
5.1. LIMITAÇÕES 81
5.2. TRABALHOS FUTUROS 81
6. REFERÊNCIAS 84

1. INTRODUÇÃO

Este capítulo exibe um panorama geral sobre o presente documento de dissertação com as seguintes seções: A seção 1.1 realiza a contextualização do tema da pesquisa; na seção 1.2 é apresentada a motivação que levou ao desenvolvimento do trabalho; a seção 1.3 descreve o objetivo geral e específicos do trabalho; na seção 1.4 é retratada a metodologia que direcionou o desenvolvimento da pesquisa; e, finalmente, seção 1.5 apresenta como o restante do documento está organizado.

1.1. CONTEXTO DE PESQUISA

A tecnologia da informação está presente em todas as esferas da sociedade e no setor da saúde tanto pública quanto privada não é diferente. Um estudo realizado pelo Comitê Gestor da Internet (CGI), no ano de 2013, detectou que no Brasil mais de 60% da população nacional procura por informações sobre saúde na internet. O acesso a essas informações permite ao paciente mais entendimento sobre sua saúde. Murray *et al.* (2005) afirmam que a internet possibilita a troca de experiências entre pacientes com problemas semelhantes, o que facilita o debate entre especialistas e enfermos, além de ser apontada como uma eficiente estratégia para gerenciar condições clínicas, oferecendo melhorias na qualidade de vida e promovendo autoconfiança, pro-atividade e autonomia aos pacientes.

Devido ao aumento contínuo dos dados na *web*, a internet transformou-se em um repositório para todo tipo de informação. Ferguson e Frydman (2004) concordam que para muitos pacientes a internet é vista como uma base de informação eficaz e de esclarecimento sobre questões ligadas à sua saúde. Atualmente, a internet é bem aceita e frequentemente utilizada, por pacientes e profissionais como fonte de informação sobre saúde. Diversos assuntos relacionados à saúde podem ser obtidos pela Internet, sendo os temas mais pesquisados dieta, ginástica, medicamentos, tratamentos experimentais e busca por médicos e hospitais (Fox, 2006).

Os autores Wietzel *et al.* (2010) alertam que a *web* é uma fonte de busca onde as pessoas procuram informações sobre cuidados em saúde mas que é aberta a vários tipos de publicação e provedores de informação, assim a qualidade das informações em saúde que são publicadas são altamente variáveis e dinâmicas. Beruel (2008) corrobora com a ideia justificando que, devido à ausência na identificação de padrões de qualidade, a internet pode representar um grande risco na área da saúde tanto entre os profissionais de

saúde como entre os pacientes. Uma pesquisa realizada pelo instituto *Britsh United Provident Association* (BUPA) no ano de 2011 ilustrou que 86% dos brasileiros com acesso à internet pesquisam informações sobre saúde, porém somente 25% confere a fonte. McDaid e Park (2010) afirmam que as novas tecnologias vêm ajudando as pessoas a conhecerem mais sobre a própria saúde e tomarem decisões com mais confiança, entretanto, é preciso se certificar de que as informações encontradas são corretas.

Em virtude da imensidão de dados contidos na internet, ferramentas que vasculham por informações sobre saúde na *web* surgem como aliadas a pacientes e profissionais no que diz respeito à manutenção da saúde. Dentre os vários tipos de aplicações computacionais existentes, destacam-se os motores de busca. Segundo Trivedi (2009), um motor de busca é um sistema de recuperação de informações projetado para ajudar a encontrar informações armazenadas em repositórios, como a *World Wide Web*, uma rede proprietária, ou até mesmo um computador particular. O motor de busca permite que um usuário solicite conteúdo que atenda a critérios específicos e recupere uma lista de itens que correspondem a esses critérios.

É nesse cenário de busca por informações sobre saúde em fontes confiáveis na internet que se desenvolveu o presente trabalho. O trabalho apresenta o desenvolvimento de um motor de busca para integrar informações de bases de dados médicas a aplicações computacionais voltadas para saúde.

1.2. MOTIVAÇÃO

Perante o quadro apresentado anteriormente, fica evidente a necessidade da criação de ferramentas que sejam alimentadas por fontes de dados especializadas, garantindo desta maneira padrões de qualidade das informações que chegam aos pacientes pela internet. Ainda nesse contexto, Soares (2004) apresenta em seu trabalho a carência de pesquisas no que diz respeito à utilização da internet para saúde dentro da realidade brasileira.

Assim, a motivação desse trabalho dá-se pela necessidade da criação de uma ferramenta que forneça dados corretos para aplicativos computacionais no âmbito da saúde. Possibilitando desta forma, que essas aplicações sejam capazes de disponibilizar conteúdos confiáveis e de qualidade, visto que as informações são advindas de bases de dados que são mantidas por profissionais da área da saúde.

A complexidade do trabalho pode ser identificada pelo desenvolvimento de uma ferramenta que seja capaz de encontrar fontes de dados confiáveis e interpretar as informações disponibilizadas pelas mesmas. Além disso, a ferramenta deve ser capaz de armazenar a grande massa de dados encontrados, filtrar e fornecer de maneira tratada as informações solicitadas pelas aplicações a qual ela está acoplada.

1.3. OBJETIVOS

O objetivo geral deste trabalho é desenvolver um motor de busca para integrar conteúdo de bases de dados médicas a aplicações de saúde, a fim de garantir conteúdo de qualidade e principalmente com alto índice de confiabilidade a essas aplicações.

1.3.1. OBJETIVOS ESPECÍFICOS

Para alcançar o objetivo geral é necessário primeiramente atingir os objetivos específicos, sendo eles:

- Desenvolver os componentes da arquitetura do motor de busca;
- Recuperar conteúdos de bases de dados médicas na *web*;
- Integrar o motor de busca a uma aplicação de saúde;
- Validar os conteúdos recuperados pelo motor.

1.4. METODOLOGIA

Metodologias de pesquisas são utilizadas para alcançar os objetivos previamente estabelecidos. Para o desenvolvimento de qualquer tipo de trabalho cientifico, faz se necessário a utilização de paradigmas ou metodologias de pesquisa para direcionamento do pesquisador. Para atingir os objetivos apresentados, o presente trabalho executou as seguintes etapas:

1. **Pesquisa Exploratória:** Inicialmente foi realizada uma pesquisa exploratória, como forma de caracterizar o problema de consumo de informações sem fundamentação cientifica. O objetivo dessa pesquisa foi realizar um estudo preliminar como forma de familiarização do contexto abordado e oferecer prosseguimento a estudos com maior precisão e profundidade;

2. **Pesquisa Teórica:** Posteriormente, foi desenvolvida uma pesquisa teórica para esclarecimento dos termos e conceitos referentes ao tema abordado no trabalho. Por meio da análise da literatura publicada, foi possível identificar os componentes que formam a arquitetura do motor de busca. Também foram identificadas bases de dados de saúde na web;

3. **Prototipação:** Embasado pelo conhecimento adquirido na revisão bibliográfica, foi dado início a etapa de prototipação do motor de busca. Foram construídos os componentes e desenvolvidas as funcionalidades de captação e disponibilização dos conteúdos. Foi especificada uma ontologia para leitura e classificação dos conteúdos recuperados;

4. **Testes Experimentais:** Finalizada a prototipação, foi iniciada a fase de testes do motor. O motor foi integrado a uma aplicação que recomenda conteúdos sobre diabetes. O objetivo da integração foi analisar se o motor era capaz de fornecer conteúdos que eram relacionados ao escopo da aplicação;

5. **Validação dos Resultados:** Por fim, foi realizada uma validação prática dos conteúdos disponibilizados pelo motor. Foi aplicado um questionário a pacientes e especialistas, com perguntas referentes aos conteúdos disponibilizados. O intuito da validação foi medir o nível de satisfação dos usuários sobre os conteúdos.

1.5. ESTRUTURA DO DOCUMENTO

O presente trabalho está organizado conforme a seguinte estrutura: No capítulo 2 a revisão bibliográfica apresenta conceitos importantes para compreensão do trabalho. A prototipação do motor de busca é ilustrada no capítulo 3, descrevendo sua estrutura e funcionalidades. No capítulo 4 são exibidos os resultados obtidos por meio de testes experimentais e também as validações dos mesmos. E, por fim, o capitulo 5 apresenta as conclusões, limitações e propostas futuras para o trabalho.

2. REVISÃO BIBLIOGRÁFICA

Uma revisão de literatura é considerada um ponto de partida para a qualquer tipo de investigação científica. O presente capítulo esclarece termos e conceitos relacionados com o tema em estudo.

2.1. FONTES DE DADOS DE SAÚDE

Conforme apresentado no guia de saúde da Bireme[1], uma fonte de informação é qualquer recurso que responda uma demanda por parte dos usuários, incluindo produtos e serviços de informação, pessoas ou rede de pessoas, programas de computador. Esses recursos, são organizados em 2 categorias principais: (GIANOTTI, PELLEGRINO e WADA, 2013)

Fontes Primárias são consideradas documentos do ponto de vista da procedência, com informação nova ou original, advindas de artigos científicos. O autor Veja (1995) coloca além dos artigos, capítulos de livros, monografias e publicações seriadas no todo ou partes como os capítulos. Nesse mesmo grupo também são considerados os documentos que incorporam informações produzidas em todos os níveis governamentais, acadêmicos, negócios e indústria.

As fontes secundárias são índices, bases de dados bibliográficas e diretórios de pesquisadores. Incluem-se também os serviços de comutação bibliográfica. De acordo com Rossi (2004), dados secundários são por definição, dados já publicados anteriormente que não foram coletados em prol da pesquisa em questão, mas que estão disponíveis para consultas. Alguns exemplos de fontes confiáveis bastante utilizados neste tipo específico de levantamento são o MELINDE e o WEBMD. Os levantamentos em fontes secundárias compreendem: levantamentos bibliográficos, documentais, estatísticos e de pesquisas previamente realizadas tanto do meio externo quanto do meio interno da empresa.

2.1.1. BASES DE DADOS DE SAÚDE

A seguir, são apresentadas e descritas algumas das bases de dados encontradas por advento da presente revisão bibliográfica.

Health Vault é um serviço online que permite a seus usuários coletar, armazenar, e compartilhar informações sobre saúde. Mantido pela empresa Microsoft, o serviço conta com uma base de dados com informações sobre saúde pessoal, clinica, fornecidas por pacientes, médicos especialistas e profissionais de saúde (HealthVault, 2018). O Health Vault pode ser acessado pelo endereço: https://international.healthvault.com/

Informatics for Integrating Biology and the Bedside é um centro de pesquisa financiado pelo *National Institutes of Health* (NIH) para fomentar o desenvolvimento da infraestrutura computacional para informática biomédica nos Estados Unidos da América. A base de dados é constituída por registros eletrônicos de saúde, resultados laboratoriais, dados genéticos e dados de pesquisas voltadas para área médica (i2b2, 2018). O i2b2 pode ser acessado pelo endereço: https://www.i2b2.org/

Medline é uma base de dados online que oferece acesso gratuito a referências e resumos de revistas científicas da área Biomédica. São indexados nesta base aproximadamente 5.400 periódicos dos Estados Unidos e de mais 80 países (Medline, 2018). A base de dados pode ser acessada pelo endereço: https://www.nlm.nih.gov/bsd/medline.html

Merck Index é uma enciclopédia farmacêutica de informações sobre produtos químicos, drogas e produtos biológicos. Construído a mais de 120 anos, o Merck contém mais de 11.500 arquivos, distribuídos em monografias, textos históricos e artigos científicos (Merck, 2018). Disponível em versão impressa e digital, a enciclopédia pode ser acessada pelo endereço: https://www.rsc.org/merck-index

O WEBMD é um site especializado em notícias sobre saúde e bem-estar humano na web. Além de disponibilizar conteúdos, o site mantém uma comunidade de pacientes, médicos especialistas e profissionais de saúde, onde existe troca de informação e conhecimento entre seus membros. Dentre seus conteúdos, são destacados: imagens médicas, prontuários de pacientes, registros eletrônicos de saúde e vídeos explicativos (WEBMD, 2018). A site pode ser acessado pelo endereço: https://www.webmd.com/

2.2. METADADOS

Para o autor Grácio (2002) o conceito de metadados diverge de acordo com o profissional e a área em que é utilizado, porém possui sempre um objetivo principal: a descrição da informação para sua busca e recuperação.

Metadados são descrições de dados armazenados em banco de dados, ou como é comumente definido "dados sobre dados a partir de um dicionário digital de dados" (DE SOUZA, CATARINO e DOS SANTOS, 2012). Neste sentido, esses dicionários normalmente são utilizados para organizar os metadados contendo seções descrevendo, numa visão geral, como os dados são subdivididos em arquivos, que campos de registros se relacionam e possuem tópicos tais como: convenções adotadas em sua definição (RIBEIRO, 1995).

A finalidade principal dos metadados é documentar e organizar de forma estruturada os dados das organizações com o objetivo de minimizar duplicação de esforços e facilitar a manutenção dos dados. Os metadados surgiram em função da necessidade de as organizações conhecerem melhor os dados que elas mantêm e conhecer com mais detalhes os dados de outras organizações. A catalogação dos dados proporciona a maior utilização deles por usuários com múltiplos interesses. Os dados precisam conter informações que auxiliem seus usuários a tomar decisões sobre a sua devida aplicação. O objetivo dessa forma de descrição documentária é colaborar na orientação, no desenvolvimento e descrição dos documentos eletrônicos, emergindo padrões, produção e manipulação da descrição por metadados (IKEMATU, 2002).

2.2.1. SGML

O SGML (*Standard Generalized Markup Language*) é um padrão internacional independente de hardware e software para a definição de métodos de representação de textos em formato eletrônico. Foi criado no ano de 1960, desenvolvido para a representação e intercâmbio de documentos, e permite que documentos armazenados eletronicamente sejam definidos conforme seu conteúdo e sua estrutura, independentemente de sua forma de apresentação (BROWN, 1996).

Segundo Goldfard (1995), o SGML é uma metalinguagem para descrição da estrutura lógica de um documento utilizando-se de marcações (*markup*). Assim, um documento SGML é uma sequência de caracteres, em formato legível, consistindo do texto do documento intercalado com comandos de marcação que identificam o início e o fim de cada item lógico. O SGML é fundamentado por três conceitos: elementos, atributos e entidade.

- **Elementos:** Os elementos são delimitados pelas *tags*. As *tags*, definem o início e o fim do texto marcado como elemento. Por exemplo: <parágrafo> aqui o texto é iniciado </parágrafo>. Pode-se também embutir elementos dentro de outros, por exemplo: <tópico> <parágrafo> aqui o texto é iniciado</parágrafo> </tópico>. Assim esse paradigma permite tratar cada unidade de informação como um objeto (ou entidade) ao qual se pode atribuir características específicas, o que possibilita maior estruturação da informação (MORENO e BRASCHER, 2007).
- **Atributos:** Proporcionam informação extra sobre o elemento que está sendo especificado. Atributos são similares a parâmetros em linguagens de programação. É sempre incluído no tag inicial de um elemento, usando a sintaxe: <parágrafo nome_do_ atributo="valor"> (ARYA e DWIVEDI, 2015).
- **Entidades:** uma entidade é um nome associado a alguma parte do documento (ou de um outro documento). A "referência a uma entidade" ("entity reference") obedece à notação: &nome-da-entidade; Entidades podem ser "internas" ou "externas" ao documento. São utilizadas para referenciar, um conteúdo ou um para um ficheiro externo onde está esse conteúdo (GUIMARÃES, 2003).

Ainda sobre elementos, um documento SGML é composto por três partes: Declaração SGML, Declaração de tipo de documento (DTD) e Instância do documento.

- Declaração SGML: Especifica conjuntos de caracteres, a sintaxe concreta, requisitos de processamento e caraterísticas opcionais contidos no documento SGML. Para Lampesberger (2016) a declaração SGML inclui informação sobre o uso de conjuntos de caracteres, a sintaxe concreta, requisitos de processamento e caraterísticas opcionais. É obrigatória a utilização da declaração para todo documento SGML. A declaração SGML e o DTD trabalham juntos;
- Declaração de Tipo de Documento (DTD): É utilizada para delimitação das marcações no código. Para Bax (2001) O DTD é uma espécie de gramática que define como as marcas devem ser interpretadas, quais as regras que restringem o uso de cada marca nos diferentes contextos do documento e, até mesmo, quando relevante for, a ordem em que as marcas devem aparecer no documento. Figura 01 mostra uma Declaração de Tipo de Documento.

Figura 01 – Exemplo de Código SGML

```
<!DOCTYPE livro
    [
        <!ELEMENT livro - - (ti, au, dp, ed, ln)>
            <!--
                COMPONENTES DO LIVRO EM FORMATO DE TAGS(MARKUPS):
                    ti = Titulo do Livro
                    au = Autor do Livro
                    dp = Data da Publicacao
                    ed = Editora
                    ln = Idioma da Publição
            -->
        <!ELEMENT (ti|au|dp|ed|ln) - (#PCDATA)>
    ]
>
```

Fonte: Autoria Própria

- Instância do documento: A instância do documento representa o documento propriamente dito que é ser marcado de acordo com as regras estabelecidas da declaração SGML e no DTD. A Figura 02 dá um exemplo de uma instância de documento SGML juntamente com a declaração DTD.

Figura 02 – Exemplo de Instância SGML

```
<!DOCTYPE LIVRO SYSTEM “livro.dtd”>
    <livro>
        <ti> A Cabana </ti>
        <au> William Paul Young </au>
        <dp> 01/07/2011 </dp>
        <ed> Windblown Media </ed>
        <ln> Inglês </ln>
    </livro>
>
```

Fonte: Autoria Própria

2.2.2. HTML

Construída utilizando conceitos da metalinguagem SGML, o HTML (*HyperText Markup Language*) é atualmente a linguagem mais popular da *web*. Foi criada no ano de 1990 por Tim Berners-Lee. Desde sua versão 1.0 até a atual 5.2, passou por diversas atualizações recebendo correções e acréscimos de recursos. Por ser baseada em SGML, faz utilização do mesmo mecanismo de marcações (*tag*).

De acordo com o autor Marcondes (2012), o HTML é uma linguagem de marcação projetada para apresentação de conteúdo. Desenvolvida para a ilustração de

elementos pré-definidos pela linguagem, assim, não sendo possível a criação de novas tags. Orientada para exibição de informações em telas.

Documentos HTML têm dois componentes principais: o cabeçalho (definido pela marca *head*) e o corpo (definido pela marca *body*). Ambos, constituem-se de várias outras marcas (ou *tags*) diferentes que dizem ao navegador como o documento deve ser apresentado na tela do computador. As marcas HTML fornecem ao navegador informações gerais sobre, organização e apresentação do texto, não se preocupando com a semântica da informação. São exemplos formatação de formatação de texto: tamanho da fonte, negrito, parágrafos, linhas horizontais, quadros entre outros (BAX, 1998). Na Figura 03, é ilustrado um trecho de código HTML.

Figura 03 – Exemplo de código HTML

```
<!DOCTYPE html>
<html>
    <head>
        <title> Título do Documento </title>
    </head>
    <body>
        <div>
            <p> Tag para criar-se uma 'caixa', que pode receber textos e imagens
            <a href="http://www.wikipedia.org">Wikipedia, A Enciclopédia Livre</a>
        </div>
    </body>
</html>
```

Fonte: Autoria Própria

A *tag* <html>: define o início do documento e indica que todo conteúdo posterior deve ser tratado como uma série de códigos HTML. <*head*>: define o cabeçalho e traz informações sobre o documento. <*title*>: define o título da página, que é exibido na barra de título dos navegadores <*body*>: define o conteúdo principal, o corpo do documento. Esta é a parte do documento HTML que é exibida no navegador. <a>: cria um link para um outro local, seja uma página, um e-mail ou outro serviço.

2.2.3. XML

A linguagem XML (*eXtended Markup Language*) foi criada em 1966 por Jon Bosak, na organização W3C. Assim como HTML, XML foi definida como um padrão de marcação para ser utilizado na Internet, constituindo-se uma versão simplificada da SGML. O objetivo principal do XML é fornecer aos desenvolvedores uma maneira de

definir e criar seus próprios marcadores e atributos em vez de estarem restritos ao esquema de marcação HTML (BAX, 2001).

Quanto ao tipo de armazenamento, os documentos XML são, armazenados geralmente em arquivos de texto com a extensão .txt. Assim, qualquer editor de texto pode ser utilizado para criar/manipular arquivos XML.

A linguagem XML pode ser vista como um conjunto de regras para a definição de marcadores semânticos, que dividem um documento em partes identificáveis. Harold (1999) destaca que o XML é uma metalinguagem que define uma sintaxe para ser utilizada na criação de outras linguagens de marcação para um domínio especifico, com estrutura e semântica próprias.

A principal vantagem do XML em relação ao HTML é a capacidade de personalização do código. Enquanto a HTML apenas trata de especificar a formatação de uma palavra ou um trecho de texto, a XML trata de criar estruturas para representar seu significado. Em outras palavras, o HTML indica como algo deve ser exibido, enquanto o XML procura indicar o que a informação significa. O próprio significado da XML sugere essa característica, pois é uma linguagem de marcação extensível (FURGERI, 2006).

O XML é um padrão para a formatação de dados, ou seja, uma maneira de organizar informações. Os documentos XML podem ser facilmente compreendidos por programadores facilitando o desenvolvimento de aplicativos compatíveis. Dá mesma forma que o HTML, o XML possui *tags*, elementos e atributos para especificação dos documentos (ALMEIDA, 2002).

Um elemento XML é composto por uma *tag* e seu conteúdo. Ele pode possuir atributos e também pode ser vazio, ou seja, não possuir nenhum conteúdo. Todo elemento possui uma *tag* de início e uma *tag* de fim, mesmo que não possua conteúdo algum. A *tag* de fim sempre será igual a *tag* de início acrescida de uma barra (/) antes do nome da *tag*. Por exemplo: <NOME_*TAG*> VALOR DA *TAG* </NOME_*TAG*>. Um atributo é colocado obrigatoriamente dentro da *tag* de início, logo após o nome do elemento. Um elemento pode conter mais de um atributo e os atributos podem ser obrigatórios ou opcionais. Geralmente um atributo é usado para qualificar ou complementar a informação marcada pelo elemento ao qual pertence (NARDON, 2000).

A seguir, é apresentado um exemplo de código XML que representa uma lista de músicas. Na Figura 04, é possível visualizar exemplos de elementos, *tags* e atributos.

Figura 04 – Exemplo de código XML

```xml
<?xml version="1.0" encoding="UTF-8"?>
<lista_de_musicas>
    <musica genero="ROCK">
        <artista>RAUL SEIXAS</artista>
        <titulo>GUITA</titulo>
    </musica>

    <musica genero="MPB">
        <artista>CAETANO VELOSO</artista>
        <titulo>SOZINHO</titulo>
    </musica>

    <musica genero="SAMBA">
        <artista>MARTINHO DA VILA</artista>
        <titulo>MULHERES</titulo>
    </musica>
</lista_de_musicas
```

Fonte: Autoria Própria

2.2.4. RDF

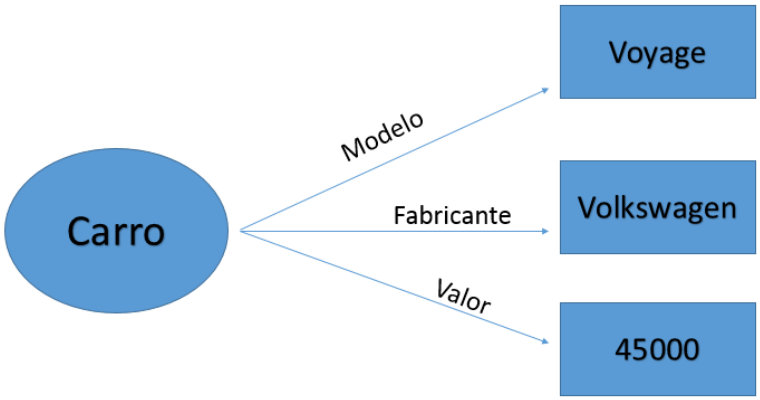
O RDF (*Resource Description Framework*) é uma linguagem para representação de informação na *web*. Trata-se de uma infraestrutura que fornece a habilidade para codificação, troca e reutilização de metadados. RDF define um modelo de dados para descrição de semântica de dados para o entendimento pelo computador. É o fundamento para o processamento de metadados (ROSA, 2002).

Para Daconta (2003), o RDF é uma linguagem construída sobre o XML e utilizada para fazer afirmações sobre entidades e documentos como um todo. Enquanto o XML serve para adicionar metadados a partes de um documento, com o RDF pode-se associar metadados ao documento como uma entidade única.

Um recurso descrito por RDF consiste em um conjunto de triplas. Cada tripla pode ser interpretada como uma afirmação sobre um recurso. A tripla possui três elementos, e a posição dos elementos define a semântica da afirmação. O formato utilizado é: sujeito, predicado e objeto. Na posição do sujeito, está o recurso sobre qual a afirmação está sendo feita. O predicado é outro recurso que denota uma propriedade do sujeito e o relaciona através dessa propriedade com o objeto. O objeto pode por sua vez, pode ser outro recurso ou possuir um valor literal, como um número ou uma sequência de caracteres (COHEN, 2010).

Um exemplo de uma tripla é ilustrado na Figura 05 em forma de grafo dirigido. No exemplo é apresentado um sujeito (carro), que possui três predicados (Modelo, Fabricante e Valor) representado pelos respectivos objetos (Voyage, Volkswagen e 45000).

Figura 05 – Exemplo de Tripla



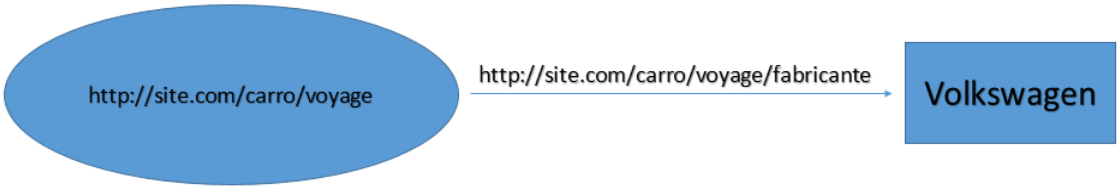
Fonte: Autoria Própria

Para representar os recursos a linguagem RDF faz uso de URI (*Uniform Resource Identifier*) de forma traduzida, identificador uniforme de recurso. O objetivo das URI'S é identificar cada um dos recursos e cada uma das propriedades de forma única e universal, de modo a ter uma semântica global e não apenas particular, algo que seja compreendido não apenas dentro de uma empresa ou organização (SAYÃO, 2007).

URIS são uma forma mais abrangente de URL (*Uniform Resource Locator*), pois não estão necessariamente ligadas à localização do recurso. Elas têm o mesmo formato de uma URL, mas são utilizadas com o intuito de identificar as coisas, enquanto uma URL identifica um endereço para a recuperação de uma informação, um documento. Uma pessoa, por exemplo, pode ser identificada por uma URI (LAUFER, 2015).

Na figura 06, pode ser observado uma descrição do sujeito (Voyage) e predicado (Fabricante), utilizando a URI http://site.com/carro.

Figura 06 – Descrição Sujeito



Fonte: Autoria Própria

Uma parte fundamental para descrição dos recursos por meio do RDF é a utilização de vocabulários para nomeação dos predicados, com intuito de estabelecer padrões para descrever as informações. O RDF permite a criação de vocabulários próprios ou, a utilização de vocabulários de terceiros que possuam uma semântica bem definida.

De forma resumida, o objetivo da utilização de vocabulários é fazer com que o significado pretendido pelo publicador dos dados seja o mesmo que o significado entendido pelo consumidor (SILVA, 2008). Na figura 07, é ilustrado um recurso descrito em linguagem XML utilizando RDF. Os dados são descritos pelo vocabulário http://site.com/carro.

Figura 07 – Exemplo de código XML

```
<?xml version="1.0"?>

<rdf:RDF xmlns:carro="http://site.com/carro#">

    <rdf:Description
        rdf:about="http://site.com/carro/Voyage">
        <carro:fabricante>Volkswagen</carro:fabricante>
        <carro:valor>45000</carro:valor>
        <carro:ano>2018</carro:ano>
    </rdf:Description>

    <rdf:Description
        rdf:about="http://site.com/carro/Fiesta">
        <carro:fabricante>Ford</carro:fabricante>
        <carro:valor>55000</carro:valor>
        <carro:ano>2018</carro:ano>
    </rdf:Description>
</rdf:RDF>
```

Fonte: Autoria Própria

A primeira linha do documento RDF é a declaração XML. A declaração XML é seguida pelo elemento raiz de documentos RDF: <rdf: RDF>. A segunda linha do código especifica que os elementos com o prefixo (carro) são referentes ao vocabulário http://site.com/carro. O elemento <rdf: Description> contém a descrição do recurso

identificado pelo atributo rdf: *about*. Os elementos: <carro: fabricante>, <carro: valor>, <carro: ano>, são propriedades do recurso.

2.2.5. PADRÕES DE METADADOS

Ainda sobre metadados, (GARCIA, 1999) afirma que a utilização de metadados permite também estabelecer padrões de dados diante da heterogeneidade das informações disponíveis em rede, principalmente as da internet. Esses padrões são organizados em estruturas formais por um conjunto de elementos para um fim específico, com a finalidade de uma melhor descrição dos recursos. (FORMENTON et al. 2017).

A seguir, apresenta-se uma breve descrição de alguns dos principais padrões de descrição de metadados (DA CUNHA, 2018).

- **MARC 21:** O MARC é um padrão para entrada e manuseio de informações bibliográficas em computadores com um protocolo de intercâmbio de dados para exportação e importação de dados. O padrão é opera como uma estrutura legível por máquina, e consegue suportar a descrição de recursos informacionais, cujo conteúdo dos seus campos é determinado por regras de catalogação;

- **MODS:** *Metadata Object Description Schema*, é um esquema de um conjunto de elementos bibliográficos que podem ser usados com diversas finalidades, e particularmente para aplicações em bibliotecas;

- **MIX:** *Metadata for Digital Still Images Standards Committee*, é um padrão que fornece um formato para intercâmbio e/ou armazenamento dos dados especificados no Dicionário de Dados - Metadados Técnico de fotografias digitais (ANSI / NISO Z39.87- 2006). Esse esquema é atualmente conhecido como "NISO Metadados para imagens em XML (NISO MIX)". MIX se expressa através da linguagem XML;

- **CORE VRA:** É um padrão de metadados para a descrição das obras da cultura visual, bem como as imagens de documentação. O padrão é

organizado pela Rede de Desenvolvimento e MARC Standards Gabinete da Library of Congress (LC) em parceria com a Associação de Recursos Visual;

- **XMP:** *Extensible Metadata Platform*, é um padrão que permite incorporar metadados dentro dos próprios arquivos mesmos durante o processo de criação do conteúdo. Com uma aplicação XMP ativada, pode-se obter informações significativas sobre um projeto (como títulos e descrições, palavras-chave e informações atualizadas de autor e direito autoral) em um formato de compreensão fácil tanto pela equipe de trabalho, quanto pelos *softwares*, dispositivos físicos e mesmo formatos de arquivo;

- **PLUS:** *Picture Licensing Universal System*, é um padrão São metadados que expressam os direitos de utilização e licenças para imagens. Fornece um conjunto integrado de padrões para informar direitos autorais e informações de propriedade que estão associadas com imagens existentes e licenciadas;

- **Dublin Core:** Este padrão é um conjunto de 15 elementos com o objetivo de descrever um recurso eletrônico. O padrão se caracteriza pela simplicidade, interoperabilidade semântica, consenso internacional, extensibilidade e flexibilidade.

2.2.6. PADRÃO DUBLIN CORE

O Dublin Core (DC), é um padrão de metadados projetado com proposito de organizar as informações nas páginas da *web*, de maneira estabelecer padrões de catalogação e classificação das informações no meio eletrônico. O DC, tem suas origens em Chicago, na 2ª Conferência Internacional sobre a WWW em outubro de 1994, quando Yuri Rubinsky, Stuart Weibel e Eric Miller integrantes da OCLC – Online Computer Library Center e Joe Hardin da NCSA – *National Center for Supercomputing Applications*, iniciaram uma discussão em semântica e WEB (PEREIRA et al 2000).

O DC é um formato menos estruturado e mais flexível, que adota a sintaxe do RDF. Estabelecido pelo Consórcio W3C, responsável pelo gerenciamento da Internet, propicia um conjunto de 15 elementos padrão, permitindo a inclusão de elementos adicionais para atender às particularidades de cada usuário. As principais características do padrão DC são: a simplicidade na descrição dos recursos, entendimento semântico universal dos elementos, escopo internacional e a capacidade de adaptação às necessidades adicionais de descrição por meio da extensibilidade (SOUZA et al. 2000).

Na tabela I, são apresentados os 15 elementos que compõem o padrão DC e uma breve especificação de cada um.

Tabela I – Elementos e Especificações do Padrão Dublin Core

Elemento	**Especificação**
Título	Nome atribuído ao conteúdo
Autor	Pessoas/organizações criadores do conteúdo
Palavras-Chave	Palavras-chave que descrevam o assunto abordado no recurso.
Descrição	Uma descrição textual(Abstract) do conteúdo.
Publicador	A entidade responsável por tornar o recurso disponível.
Colaborador	Pessoa/organização que colaborou para criação do conteúdo.
Data	A data em que o recurso tornou-se disponível.
Tipo	A categoria do recurso, como texto, imagem, som, dados etc.
Formato	O formato do dado do recurso, usado para identificação.
Identificador	Identificador único, por exemplo: URL ou URI
Fonte	Espaço onde o conteúdo foi disponibilizado.
Idioma	O idioma em que é descrito o conteúdo.
Relação	Especifica como o conteúdo se relaciona com a fonte
Cobertura	Localização Geográfica de onde está armazenado o conteúdo
Direitos Autorais	Uma declaração de direito sobre a propriedade intelectual.

2.3. RECUPERAÇÃO DA INFORMAÇÃO

Com a crescente expansão da web, a quantidade de dados nela disponível aumenta de maneira desenfreada. Na grande esfera da Ciência da Informação (CI), a área da Recuperação de Informação (RI) é encarregada da missão de idealizar técnicas de representação, localização e armazenamento de informações em todos os tipos de domínios do conhecimento (BAEZA-YATES, RIBEIRO-NETO, 2011).

Apesar de ser uma área com diversos trabalhos realizados, não existe uma definição exata do conceito de recuperação de informação. Uma das primeiras aparições

do termo foi apresentada por Rijsbergen (1979). O autor afirma que sistemas de recuperação de informação (SRI) são desenvolvidos para recuperar conteúdos indexados por bases de dados. Para que os conteúdos sejam recuperados, é necessário que os mesmos estejam indexados de forma que satisfação condições previamente definidas. Em um SRI, faz-se necessária uma interpretação sintática e semântica dos documentos, de forma que é os resultados das consultas sejam precisos e relevantes.

No contexto da recuperação da informação, todo dado e informação armazenado deve possuir forma e estrutura definida pela fonte que os mantém. Essas fontes atribuem características que descrevem as informações armazenadas de forma a catalogar os dados, e, disponibiliza-los para requisições em forma de consultas. As consultas são expressas em linguagem natural e são constituídas pelos termos solicitados pelos usuários. Os resultados representam uma resposta exata à consulta efetuada pelo usuário (AGNER e MORAES, 2008).

2.3.1. MODELOS DE RECUPERAÇÃO DA INFORMAÇÃO

Para localização e disponibilização dos documentos, os SRI fazem uso de modelos de recuperação de informação. São três os principais modelos: Booleano, Vetorial e Probabilístico.

Modelo Booleano

O Modelo Booleano é um modelo de recuperação que se destaca pela simplicidade e facilidade de entendimento. É baseado pela lógica booleana e teoria dos conjuntos. As consultas são constituídas pelos termos da busca e por 3 tipos de operadores lógicos: AND, OR e NOT (FERNEDA, 2012).

Os autores Silva et al (2013) reiteram que no modelo booleano é possível classificar o conteúdo em somente 2 tipos: irrelevante ou relevante. Devido a essa rigidez do modelo, não é possível encontrar um resultado parcial para cada consulta. A principal desvantagem do modelo é a incapacidade de ordenação do resultado da consulta. Dessa forma, conteúdos pouco relevantes possuem mesmo peso que conteúdos muito relevantes. As principais vantagens do modelo booleano são a simplicidade e baixo formalismo.

Em uma consulta hipotética representada por: (c = x1 AND x2), todos documentos indexados pelos termos x1 e x2 serão recuperados. Essa equação simula à intersecção do

conjunto de documentos indexados por x1 com o conjunto de documentos indexados por x2. Aplicando o operador lógico OR, é realizada a união entre o conjunto de documentos indexados pelos termos da consulta. Utilizando o operador NOT, são recuperados documentos que possuem termos que sejam diferentes aos termos da consulta.

Modelo Vetorial

O modelo vetorial diferentemente do booleano, proporciona uma recuperação de documentos que correspondam a consulta de forma parcial, e não somente exata. Isso é possível pela capacidade de atribuição de pesos, tanto para os documentos, como para as consultas. Como resultado, é retornado uma lista ordenada pelo grau de similaridade dos conteúdos com os elementos da busca (GOMES, 2009).

O modelo vetorial representa cada documento como um vetor de termos, e, cada termo possui um valor associado. Esse modelo, permite o uso de pesos não binários, associados aos termos de índice, proporcionando combinação entre a consulta e os documentos da coleção. Ribeiro-Neto e Baeza (1999) afirmam que o modelo vetorial apresenta resultados mais precisos em relação ao modelo booleano, devido a capacidade de atribuir valores pesos aos documentos e consultas.

A representação em forma de vetor, permite ao modelo determinar a semelhança utilizando a formula de cálculo do cosseno do angulo entre os termos dos documentos e a consulta. Na equação a seguir, é apresentado um exemplo de cálculo de similaridade entre documento e consulta. Em um vetor de tamanho **(n)**, a similaridade **(s)** é calculada de acordo com o peso **(p)** associado aos termos dos documentos **(d)** e a consulta **(c)**. O peso de cada termo dos documentos é representado por $(\boldsymbol{p_{i,d}})$. O peso de cada termo referente a consulta é representado por $(\boldsymbol{p_{i,c}})$.

$$S(d_j, c) = \frac{\sum_{i=1}^{n}(p_{i,j} \times p_{i,c})}{\sqrt{\sum_{i=1}^{n} p_{i,j}^2} \times \sqrt{\sum_{i=1}^{n} p_{i,c}^2}}$$

Modelo Probabilístico

O modelo probabilístico possui a premissa da classificação dos documentos segundo relevância e probabilidade de acordo com a especificação da consulta. Embasado

pela teoria das probabilidades da área da Matemática, o modelo tenta encontrar um conjunto de documentos considerado ideal para a consulta (CARDOSO, 2004). Em relação aos outros 2 modelos apresentados, o modelo probabilístico possui uma particularidade: necessita da interação do usuário (*Feedback*) para refinamento no processo de recuperação.

Os autores Ribeiro-Neto e Baeza (1999) revelam que, dado uma consulta e um documento na coleção, o modelo estima a probabilidade de o documento ser relevante para a consulta. O modelo determina que a probabilidade de relevância depende somente da consulta e da descrição do documento. Ainda, o modelo determina que existe um subconjunto de todos os documentos que o usuário pretende como resultado de sua busca para a consulta. O resultado ideal, deve maximizar toda probabilidade relevante para o usuário. Documentos contidos no conjunto são previstos ser relevante para a consulta. Documentos fora deste conjunto são classificados como não relevantes.

O objetivo deste modelo é estimar, a probabilidade do documento, dentro do conjunto de resposta ideal, ser relevante para a consulta. Por fim, é iniciada uma interação com o usuário. O objetivo da interação é aperfeiçoar a descrição probabilística do conjunto de resposta ideal. O usuário analisa os documentos recuperados e atribui um valor de interesse. O sistema usa esta informação para refinar a descrição do conjunto de resposta ideal. Repetindo-se muitas vezes este processo, existiria então uma evolução e consequentemente uma aproximação do conjunto de resposta ideal. Assim, o usuário deverá ter em mente no princípio, a necessidade, para prever o conjunto de resposta ideal (BARTH, 2013).

O processo é dividido em duas etapas. Inicialmente é calculada as probabilidades de um documento ser relevante $\boldsymbol{pr}(+\boldsymbol{R_c}|\,\boldsymbol{d})$ e não-relevante $\boldsymbol{pr}(-\boldsymbol{R_c}|\,\boldsymbol{d})$. Dessa forma, é atribuído o valor ao peso **(P)** para cada documento **(d)** em relação a consulta **(c)**.

$$P_{d|c} = \frac{pr(+R_c|\,d)}{pr(-R_c|\,d)}$$

Posteriormente, é aplicado o teorema de Bayes para medição das estimativas de relevância baseadas nos termos da consulta. São atribuídos valores binários a cada termo $(\boldsymbol{x_i})$ do documento **(d)**. Inicialmente é considerado que todos documentos possuem mesma probabilidade de relevância, logo, é atribuída a seguinte equação para os pesos, $\boldsymbol{P_{ci}} = \mathbf{log}\,\boldsymbol{a_{ci}}\,(\mathbf{1}-\boldsymbol{a_{ci}}) \div (\mathbf{1}-\boldsymbol{b_{ci}})$. Na equação temos que $\boldsymbol{a_{ci}}$ é a probabilidade de

que um termo $\boldsymbol{i}$ ocorra em um documento relevante. Já em $\boldsymbol{b_{ci}}$ temos a probabilidade de um termo $\boldsymbol{i}$ ocorra em um documento não relevante. Aplicando transformações algébricas, é calculada a similaridade do documento para com a consultada pela seguinte equação.

$$s(d,c) = P_{d|c} = \sum_{i=1}^{n} x_i \times P_{ci}$$

2.3.2. AVALIAÇÃO DA RECUPERAÇÃO DA INFORMAÇÃO

Como forma de analisar e avaliar os resultados dos modelos de recuperação, os sistemas de recuperação de informação fazem uso de métricas de avaliação. O objetivo dessa avaliação é mensurar a qualidade das respostas que são retornadas. Segundo Costa (2008) os documentos que correspondem ao resultado da consulta podem ser classificados nos seguintes conjuntos:

Figura 08 – Tipos de Documentos

Fonte: Autoria Própria

- Recuperados (Verdadeiro Positivo): Refere-se aos conteúdos que o SRI compreende que respondem a busca do usuário;

- Não-Recuperados: Refere-se aos conteúdos que o SRI compreende que não respondem a busca do usuário;

- Relevantes: Refere-se aos conteúdos que o SRI compreende que são relevantes para a busca do usuário;

- Não-Relevantes: Refere-se aos conteúdos que o SRI compreende que são irrelevantes para busca do usuário.

Para avaliação dos resultados, são aplicadas métricas quantitativas de forma a analisar a relação da busca com os documentos retornados. As métricas são relacionadas diretamente com os resultados fornecidos pelo SRI. Apesar da utilização de métricas quantitativas, com a avaliação, é possível analisar a qualidade das respostas retornadas (Buckland e Gey, 1994). O autor Lancaster (2004) destaca as principais métricas de avaliação para SRI, sendo essas: Revocação (*Recall*), Precisão (*Precision*).

A métrica Revocação (*Recall*) é utilizada para calcular a capacidade de um SRI, de recuperar todos os documentos que correspondem de maneira exata a busca realizada. Por exemplo: em uma base de dados, onde existam 20 conteúdos que correspondam aos termos de uma determinada busca, e o SRI consegue recuperar 17 desses conteúdos, temos um índice de Revocação de 85 por cento. A seguir, a equação para cálculo da Revocação.

$$Revocação = \frac{Documentos\ Relevantes\ Recuperados}{Total\ de\ Documentos\ Relevantes} \times 100$$

Já a métrica Precisão (Precision), avalia a capacidade de um SRI, de recuperar somente os documentos que são considerados relevantes dentro do intervalo dos conteúdos correspondentes a busca. A métrica procura dentro do conjunto resposta, quais são os conteúdos considerados relevantes com base em algum parâmetro pré-

estabelecido. Por exemplo: Em um conjunto de 20 conteúdos recuperados que respondem a determinada busca, a métrica precisão identifica que somente 10 são de fato relevantes. Desta forma temos um índice de precisão de 50 por cento.

$$Precisão = \frac{Total\ de\ Documentos\ Relevantes\ Recuperados}{Total\ de\ Documentos\ Recuperados} \times 100$$

Ao analisarmos as 2 métricas apresentadas, é possível identificar um cenário de relação inversamente proporcional entre as 2 abordagens. Um SRI que possui uma indexação com diversos termos variadas a revocação será alta, mas com precisão baixa, porém, ao considerar apenas conceitos importantes na indexação, haverá uma alta precisão com baixa revocação.

2.4. MOTOR DE BUSCA

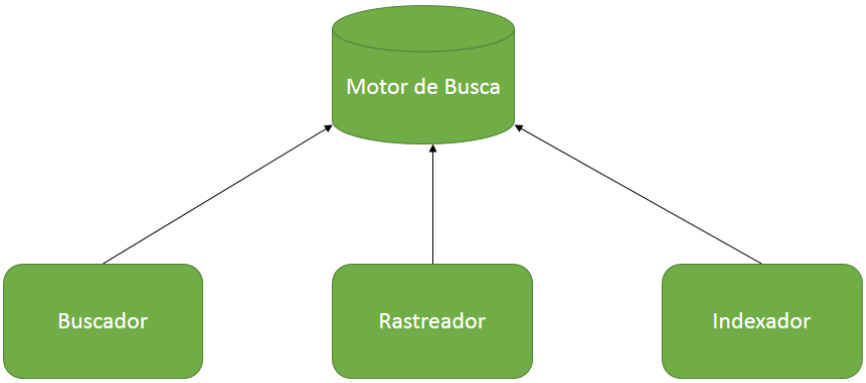
Com a disseminação de dispositivos com acesso à internet, a tarefa de fornecimento e busca por informação na web se tornou uma atividade recorrente no dia-a-dia da sociedade. Em consequência da grande quantidade de conteúdos disponíveis na web, surge o problema da falta de padronização dos documentos. As fontes que disponibilizam as informações cada dia se tornam mais dinâmicas, personalizadas, os conteúdos são atualizados frequentemente e os documentos produzidos são dos mais variados tipos. Dessa forma, justificando cada vez mais a utilização de sistema de fazem uso de técnicas de recuperação de informação. (PESSONI, 2012).

Nesse contexto, motores de busca surgem como uma alternativa para o problema de localização de informações na web. Um motor de busca é uma ferramenta embasada por técnicas de recuperação de informação, que é projetada para encontrar informações em repositórios de dados. Os motores de busca são usados por centenas de milhões de pessoas que juntos emitem centenas de milhões de consultas diariamente, para buscar informações de bilhões de páginas na Web. Embora a tecnologia de banco de dados forneça a chave para organizar e recuperar informações estruturadas, a tecnologia dos mecanismos de pesquisa é usada para indexar e consultar informações que vêm em uma infinidade de formatos de muitas fontes diferentes (CROFT, 2010).

Existem na literatura vários tipos de motores de busca, voltados para as mais diversas áreas do conhecimento. A estrutura arquitetural do motor pode variar conforme

o objetivo pelo qual o motor é construído. Na figura 08, são apresentados os componentes básicos que constituem a arquitetura de um motor de busca.

Figura 09 – Componentes do Motor de Busca

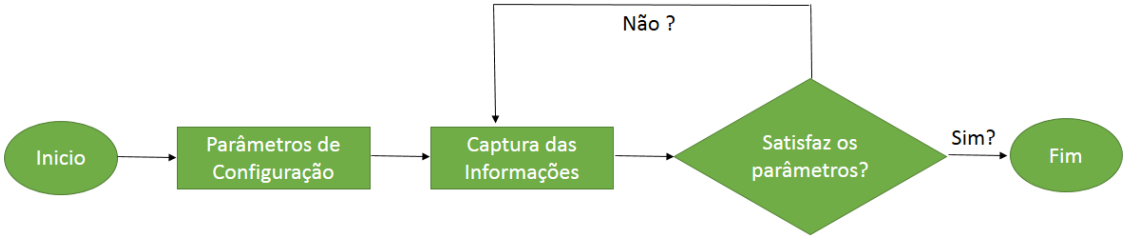


Fonte: Autoria Própria

2.4.1. RASTREADOR

O Rastreador (*crawler*), também chamado de aranha (*spider*) ou robô (*robot*) é definido como um script computacional ou software que é programado para navegar em repositórios de dados de maneira sistemática e automatizada. É o principal componente da arquitetura de um motor de busca. O Rastreador percorre todo domínio que está inserido. Domínio esse, podendo ser a web, uma base de dados ou um computador pessoal. O objetivo do Rastreador é capturar páginas web, documentos textuais e diversos tipos de arquivos, como imagens, vídeos entre outros. Os Rastreadores trabalham de maneira a criar uma cópia dos conteúdos recuperados para posteriormente serem processados pelo indexador (DIXIT e SHARMA, 2010). Na figura 09, é ilustrado o funcionamento do Rastreador.

Figura 10 – Fluxograma de Rastreamento



Fonte: Autoria Própria

O processo de rastreamento se inicia pelos parâmetros de configuração do Rastreador. No caso de um Rastreador para páginas web, o processo começa com um conjunto inicial de links conhecido como URLs iniciais. Dessas URLS, são analisadas toda sua estrutura e baixados os conteúdos e links presentes. As páginas recuperadas são armazenadas, de modo que, com a ajuda do indexador, elas possam ser recuperadas posteriormente. Esse processo é repetido até que sejam satisfeitos os parâmetros passados inicialmente. Da mesma forma, para rastreadores de conteúdos como texto, imagens e vídeos, o rastreador percorrer todo ambiente em que está inserido, capturando os conteúdos até contemplar os parâmetros configurados (GARG, GUPTA e SING, 2017).

A maneira como as informações já armazenadas são modificadas varia conforme a programação do motor de busca. A enorme velocidade com que novas informações são disponibilizadas na web, faz com que o motor de busca precise verificar se sua base de dados encontra-se atualizada. Assim, os motores de busca de maneira periódica, realizam varreduras pelas páginas e documentos que foram capturados anteriormente, com intuito de obter uma versão mais atual dos dados, e, até mesmo, verificar se essas informações ainda existem (SELVAKUMAR et al, 2012).

No momento da concepção dos rastreadores, é possível configurá-los para que trabalhem na captura de diversos tipos de conteúdo, assim, os rastreadores são classificados em dois tipos: Rastreador Web e Rastreador Especializado.

- Um rastreador da Web é projetado para coletar todos os tipos de páginas web que estão ao seu alcance. A partir de um determinado conjunto de URLs e seus links, é iniciado o processo de análise e armazenamento das páginas. Neste rastreador, a capacidade de busca um e armazenamento se estende por toda web. Porém, por se tratar de um ambiente de dimensão incalculável, é exigida uma enorme capacidade de processamento e recursos computacionais (DESAI et al, 2015).
- O rastreador especializado é planejado para coletar documentos relacionados a assuntos específicos. O objetivo do rastreador focalizado é procurar seletivamente por páginas e documentos apropriadas a um conjunto predefinido de assuntos. Ele rastreia apenas os documentos relevantes, o que gera economia significativa em recursos de hardware e de rede (LI, WU e LOU, 2015).

2.4.2. INDEXADOR

O indexador é o componente que armazena e organiza os conteúdos que serão utilizados para responder as consultas enviadas ao motor de busca. O processo de indexação dos documentos acontece em duas etapas.

Primeiramente, o indexador recebe os conteúdos, realiza leitura dos mesmos e os descreve com base em seus termos, criando assim, um vocabulário de termos referente a cada conteúdo. Em seguida, o indexador faz uso de alguma técnica ou algoritmo para classificar e ranquear os conteúdos com base no em cada vocabulário de termo. O indexador gera uma lista de todos documentos que são organizados por ordem de relevância (Davanzo, 2016).

O indexador cria índices para descreverem os documentos recuperados. Esses índices, também chamados de metadados, são descritos com partes de um documento que são armazenadas no indexador e usados para responder as consultas. Exemplos de termos de índice são palavras, frases, nomes de pessoas, datas e links que fazem parte do corpo do documento. O conjunto de todos os termos indexados para uma coleção de documentos é chamado de vocabulário de índice (BÜTTCHER et al 2016). Na tabela 2 é apresentado um conteúdo fictício.

Tabela 2 – Exemplo de Conteúdo

Documento	**Texto**
1	O diabetes é uma doença
2	que afeta o nível de
3	glicose no sangue.

Na tabela 3 é ilustrado o processo de indexação do conteúdo apresentado. É utilizada uma clássica estrutura de indexação chamada índice invertido. Neste exemplo, o índice é formado segundo as palavras que aparecem no texto. São contabilizadas quantas vezes cada palavra aparece nos documentos. Além disso, armazena-se o posicionamento de cada palavra dentro do documento.

Tabela 3 – Índice Invertido

Palavra-Chave	Número de Recorrências	Documento	Posição
O	2	1;2	1;3
diabetes	1	1	2
É	1	1	3
uma	1	1	4
doença	1	1	5
Que	1	2	1
Afeta	1	2	2
nível	1	2	4
de	1	2	5
glicose	1	3	1
no	1	3	2
sangue	1	3	3

2.4.3. BUSCADOR

O buscador é a interface entre o usuário e os conteúdos indexados. A consulta do usuário é traduzida de forma que o buscador possa verificar a disponibilidade dos dados no indexador. É constituído de um espaço para informar as palavras-chave e obter os resultados. Os documentos resultantes de um motor de busca, são retornados como resposta a uma pesquisa, normalmente incluem uma lista de conteúdo respectivos. O buscador exibe os títulos, um link para a página, e uma curta descrição mostrando onde as palavras chaves são correspondentes aos conteúdos dentro da página (FRITZEN, 2012).

2.5. ONTOLOGIAS

O termo ontologia é remetido a uma área da filosofia que investiga a natureza dos indivíduos. Para Gruber (1993), uma ontologia pode ser definida como uma especificação explícita de uma conceptualização, e também uma formalização dos conceitos e relacionamentos em um domínio específico.

Especificamente na Ciência da Computação, as ontologias possuem um papel de representar o conhecimento de maneira que o mesmo seja interpretado por seres humanos e computadores. Os autores Uschold e Gruninger (1996) afirmam que um dos fatores para a disseminação das ontologias é a premissa de entendimento e compartilhamento comum de algum domínio de conhecimento que possa servir como meio de comunicação e fonte de conhecimento.

No campo da Inteligência Artificial (IA), o autor Kiryakov (2006) define ontologia como quatro elementos que se correlacionam: Classes, Relações, Instâncias e Axiomas. As classes descrevem os termos e conceitos referentes a algum domínio do conhecimento. As relações, também chamadas de propriedades, realizam as ligações entre as classes do domínio. Instâncias são representações de indivíduos que possuem características relacionadas às classes ou não. E, por fim, os axiomas representam sentenças e condições que são utilizadas para restringirem o limiar de cada classe. Dentre as classificações dos tipos de ontologias destacam: ontologias de domínio e ontologias de tarefa.

Ontologias de domínio expressam caracterizações de domínios particulares, descrevendo o vocabulário, conceitos e regras de uma grande área. Martins (2009) relata que os objetivos principais de se desenvolver uma ontologia de domínio são: compartilhar informação, reusar elementos do domínio, tornar suposições do domínio explícitas, separar conhecimentos de domínio de conhecimento operacional e analisar o conhecimento do domínio. As ontologias de tarefa apresentam conceitos que podem colaborar na resolução de problemas, independente do domínio que sucedem. Isotani e Bittencourt (2015) simplificam que ontologia de domínio representa o conhecimento sobre uma área, enquanto a ontologia de tarefa representa a habilidade de aplicar esse conhecimento para resolver problemas em diferentes situações.

Com a crescente disseminação de informações, há também o crescimento exponencial da utilização das ontologias. Santos Neto (2013) relata que ontologias no âmbito da informática possuem aplicações em processamento de linguagens naturais, sistemas de aprendizagem, gestão do conhecimento, comércio eletrônico e recuperação da informação. Na esfera de ensino-aprendizagem, ontologias têm se mostrado úteis na concepção de vários tipos de ambientes educacionais, ambientes web, modelos educacionais formais, estruturação dos tutores inteligentes, aprendizagem colaborativa, dentre outros (Souza, Duran e Viera, 2014).

Em classificação de conteúdo, as ontologias atuam como repositórios de conhecimento que estipulam a relação semântica de uma solicitação e uma base de dados. Ferneda et al. (2017) concordam que as ontologias, vistas como vocabulários controlados, possibilitam um enriquecimento das representações dos documentos e das expressões de buscas, proporcionando convergência entre a linguagem do indexador e a linguagem do usuário, elevando consideravelmente a eficácia da classificação de um conteúdo.

2.5.1. METODOLOGIAS DE DESENVOLVIMENTO

Para desenvolvimento de ontologias formalizadas, faz-se necessário a utilização de algum procedimento formal que guie o processo da especificação da mesma. Dentre as diversas metodologias existentes destaca-se o Ontology Development 101 (OD101) proposto pelas autoras Noy e McGuinnes (2001). A metodologia 101 define um processo de *design* interativo, que vai sendo refinado conforme as etapas vão sendo concluídas. A seguir, são apresentados os passos da metodologia:

- **Delimitação do Domínio e Escopo:** A primeira etapa do OD101 consiste na identificação do domínio em que a ontologia irá atuar. Para isso, são criadas questões de competência, que possuem relação com a esfera de atuação da ontologia. Consistem em um conjunto de questões que representam demandas e indagações do domínio em que a ontologia está inserida. São elaboradas em linguagem natural, de modo que a ontologia, por meio de suas classes e relações, deve ser capaz de respondê-las corretamente.

- **Reutilização de Ontologias:** Nesta etapa, deve ser considerada a reutilização de ontologias que atuam na mesma esfera da ontologia proposta. Assim, é possível aproveitar, refinar e entender termos já estabelecidos em outras ontologias. Existem repositórios eletrônicos que são gratuitos e armazenam diversos tipos de ontologia. Alguns exemplos: BioPortal, OBO Fundry, Ontobee, Swoogle;

- **Definição de Termos Relevantes:** Nesta etapa, são elencados termos importantes a respeito da área de atuação da ontologia. Devem ser

considerados conceitos mais abstratos às definições mais concretas. Essa etapa é de suma importância, pois, a partir desses termos serão definidos os passos restantes do OD101;

- **Definição de Classes e Hierarquias:** Fazendo uso dos termos definidos na etapa anterior, são definidas as classes da ontologia. Existem 3 abordagens para construção da hierarquia. *Top-down*: O processo começa com a definição das classes mais gerais no domínio e posteriormente classes mais especificas; *Bottom-up*: Realiza o processo inverso. É iniciado pelas classes mais especificas, partindo para as mais gerais; Por fim o processo *Combination*: essa abordagem utiliza uma estratégia hibrida dos processos anteriores;

- **Definição de Propriedades:** Por meio da definição de propriedades, é possível para ontologia relacionar as classes entre si e também aos indivíduos que pertencem a elas;

- **Definição de Características:** Nesta etapa, são definidas as características das classes e propriedades. Desta forma, é possível restringir quais características os indivíduos precisam ter para ser considerados membros das classes;

- **Elaboração de Instâncias:** O último passo, consiste na criação de instâncias. Essas instancias possuem indivíduos que contém características referentes as classes e propriedades da ontologia.

2.6. TRABALHOS CORRELATOS

O campo de estudo em informações sobre saúde na *web*, mais especificamente o desenvolvimento de ferramentas de busca dessas informações, é uma ampla área de pesquisa com diversos trabalhos desenvolvidos, dentro os quais podemos destacar alguns.

No trabalho de Tang *et al.* (2008), é apresentado um mecanismo de pesquisa especializado na *Web* para recuperação de informações médicas. O MedSearch, como é

chamado, faz utilização do método OKAPI para ranqueamento dos documentos recuperados. O motor é capaz de lidar com consultas curtas e extensas em idioma inglês e sugere frases médicas relacionadas à consulta do usuário. Os autores afirmam que o MedSearch pode processar consultas longas a uma velocidade comparável a dos motores de busca tradicionais da *Web*.

Em Luo (2009) é descrito o desenvolvimento do iMED, um motor de busca médico inteligente. Definido pelo autor como um motor de busca vertical, o iMED vasculha alguns sites médicos selecionados e de alta qualidade em vez de todos os sites da *Web*, para evitar o distúrbio de páginas de baixa qualidade e sites irrelevantes no contexto da pesquisa. Embasado pela ontologia MeSH, as frases médicas encontradas são classificadas e recomendadas de acordo com a consulta do usuário. O iMed melhora a satisfação dos usuários ao realizar pesquisas médicas com eficiência e eficácia segundo relato do autor.

Gupta e Bhatia (2013) apresentam a implementação do rastreador HiCrawl. O rastreador localiza documentos médicos de bases de dados estruturadas e não-estruturadas. HiCrawl faz uso de técnicas de hierarquia de classificação para enriquecer a consulta do usuário com termos semelhantes ao pesquisado. As possíveis respostas para consulta são pontuadas e apresentadas de forma decrescente para visualização do usuário.

O trabalho de Loane *et al.* (2007) descreve a implementação do motor de busca médico Essie. O motor de busca utiliza técnicas de expansão de consulta e classificação de relevância das possíveis soluções. Embasado por um metatesauro da Biblioteca Nacional de Medicina dos Estados Unidos, os termos pesquisados pelo usuário são expandidos junto a sinônimos para uma maior abrangência da consulta. O algoritmo de pontuação fica encarregado de elencar as melhores páginas para a pesquisa do usuário. Como método de validação, os autores submeteram o Essie a uma conferência sobre ferramentas de recuperação de texto (TREC). O motor Essie foi classificado em primeiro lugar perante 25 concorrentes como ferramenta mais eficaz e eficiente.

O estudo realizado por Dragusin *et al.* (2013) ilustra um motor de busca para doenças raras chamado FindZebra. Baseado no sistema de recuperação de informação INDRI, o motor utiliza técnicas de indexação e recuperação para localização e interpretação dos documentos encontrados. Os autores realizaram um experimento comparativo com o motor de busca Google e concluíram que o FindZebra superou o concorrente em métricas de desempenho padronizadas e especificamente precisão na classificação das informações recuperadas.

Can e Baykal (2007) propõem um motor de busca para usuários sem experiência médica denominado MedicoPort. Desenvolvido na linguagem Java o motor faz uso de um dicionário de termos médicos para elevar o índice semântico das consultas. Os autores avaliam os índices de recuperação e precisão através de experimento comparativo com outro motor de busca. Os autores concluem que seu motor de busca é uma alternativa promissora para sistema de busca de informações médicas.

Ramirez *et al.* (2011) apresentam o motor de busca semântico MentalWatch, desenvolvido para pacientes e médicos inseridos no contexto de saúde mental. A consulta do usuário é processada por dois componentes. O primeiro é um dicionário que pode enriquecer com sinônimos a pesquisa. O segundo é definido como uma arvore de condições com conexões semânticas entre os sintomas. Como validação, os autores aplicaram um conjunto de 10 consultas com termos sobre saúde mental no MentalWatch e em 3 motores de busca. O MentalWatch obteve 65% de relevância em relação à precisão na recuperação dos termos pesquisados pelos usuários. Os autores concluem que MentalWatch fornece respostas aprimoradas as consultas dos usuários em tempo de resposta hábil. A seguir, na Tabela 4, são apresentadas algumas características dos trabalhos encontrados.

Tabela 4 – Análise de Competidores

Autor	**Integração de Dados**	**Classificação Semântica**	**Base de Dados Saúde**
Tang *et al.* (2008)	-	X	-
Luo (2009)	-	X	-
Gupta e Bhatia (2013)	-	X	X
Loane *et al.* (2007)	-	X	X
Dragusin *et al.* (2013)	-	X	X
Can e Baykal (2007)	-	X	X
Ramirez *et al.* (2011)	-	X	X

Apesar de diversos autores retratarem a utilização de motores de busca para recuperação de dados sobre saúde, é possível observar que nenhum dos trabalhos encontrados possuem o propósito de servir como um integrador de dados a aplicações de terceiros. Desta forma, o presente trabalho se diferencia dos demais, visto que o motor de busca possui uma proposta de fornecer conteúdos para outras aplicações. Um importante fator de confiabilidade do motor de busca, é o fornecimento de conteúdos com embasamento cientifico, visto que os conteúdos que são recuperados pelo motor são advindos de bases de dados médicas. Além disso, o motor se apresenta como uma ferramenta genérica, podendo ser configurado para busca de conteúdo de diversas patologias distintas.

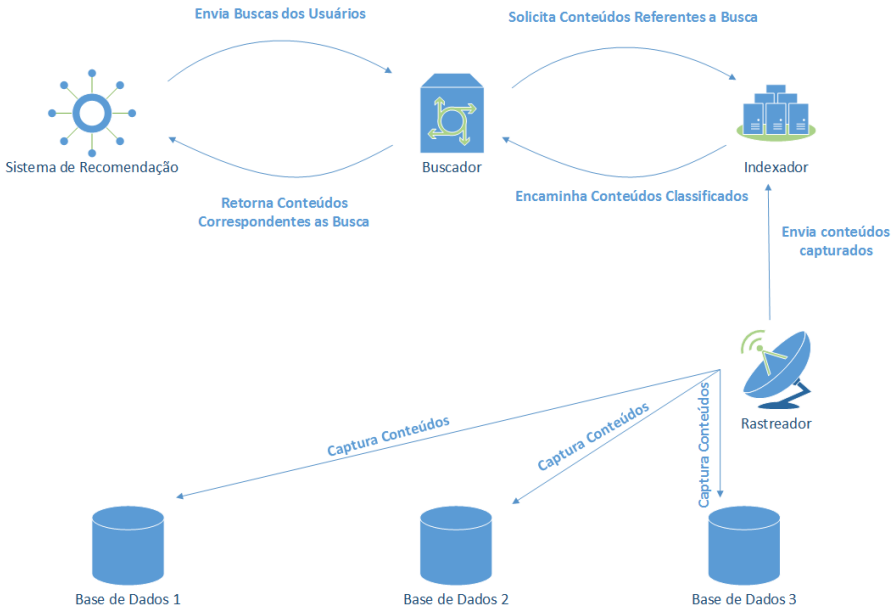
3. MOTOR DE INTEGRAÇÃO DE DADOS

O presente trabalho possui a premissa de alimentar com informações advindas de bases de dados médicas, aplicações voltadas para a área de saúde. Para isso, foi desenvolvido o protótipo do motor de busca, com enfoque em disponibilizar conteúdos médicos aos usuários dessas aplicações. Neste capítulo, são apresentados os componentes da arquitetura do motor, bem como suas respectivas funcionalidades.

3.1. ARQUITETURA DO MOTOR

O motor atua como um intermediário entre aplicações e bases de dados. O processo de comunicação entre as aplicações e o motor é baseado em protocolo HTTP. O motor possui a flexibilidade de pesquisar por conteúdos de diversas patologias diferentes. O motor recebe os parâmetros de busca das aplicações em forma de requisições, realiza a varredura nas bases de dados, e retorna para às aplicações uma lista de conteúdos que satisfaçam os parâmetros recebidos. O motor de busca foi desenvolvido em linguagem PHP e os dados são armazenado em banco de dados MySQL. Na figura 11, é ilustrada a arquitetura do motor de busca.

Figura 11 – Arquitetura Motor de Busca



Fonte: Autoria Própria

O buscador é o canal de comunicação do motor de busca. Ele é responsável por fornecer acesso ao motor de busca, cadastrar os usuários e aplicações, passar para o rastreador os parâmetros de busca e disponibilizar os conteúdos para as aplicações cadastradas. O rastreador é o principal componente do motor de busca. Ele é responsável por acessar as bases de dados, coletar os conteúdos referentes aos parâmetros, realizar uma primeira filtragem dos documentos e repassa-los ao indexador. O indexador é o componente do motor que tem como função receber os conteúdos capturados e realizar a classificação dos mesmos. Para isso, o indexador faz uso de ontologias que descrevem os termos e conceitos referentes aos parâmetros de busca e realiza a classificação dos conteúdos com bases nos metadados que os descrevem.

3.2. INTERFACE DO MOTOR

A interface de integração (buscador) é a porta de entrada para acesso ao motor de busca. Através da interface, é possível realizar o cadastro no sistema para dar início ao processo de busca pelos conteúdos. Na figura 12, é apresentada a tela inicial do motor.

Figura 12 – Tela Inicial

Fonte: Autoria Própria

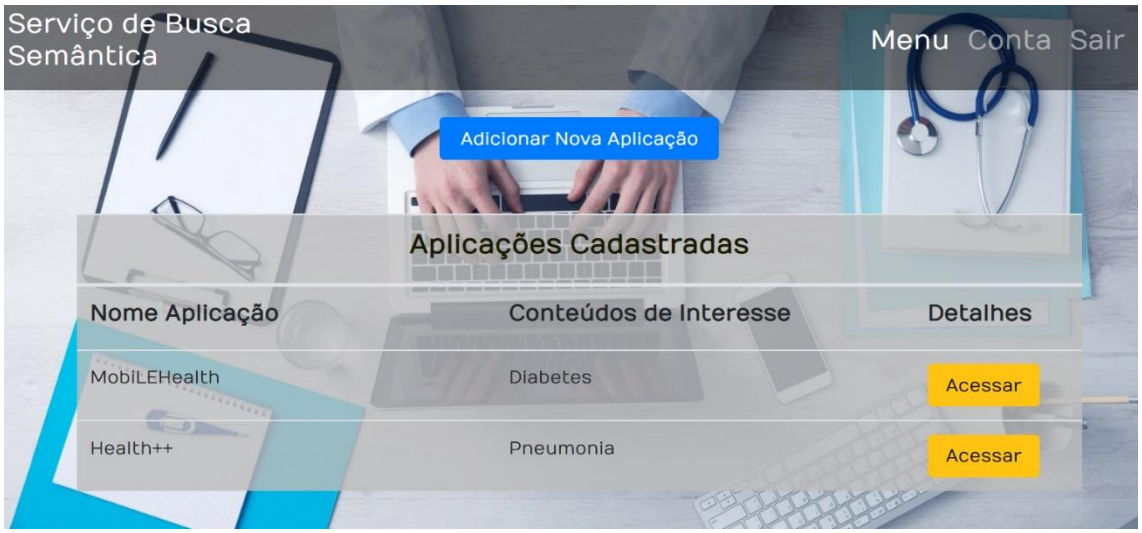
Para ter acesso às funcionalidades do motor de busca, primeiramente é necessário que o gestor da aplicação que deseja receber os conteúdos, se cadastre no motor. A seguir, a figura 13 apresenta a tela de cadastro no sistema.

Figura 13 – Tela de Cadastro

Fonte: Autoria Própria

Finalizado o cadastro, o usuário retorna para tela inicial para que possa inserir os dados nos campos login e senha, de forma que tenha acesso às funcionalidades do motor. Realizada a autenticação, o usuário é redirecionado para tela principal do motor. Na tela principal, é permitido ao usuário visualizar suas aplicações cadastradas para receber os conteúdos e, também é possível inserir novas aplicações. Na figura 14, é ilustrada a tela principal do motor.

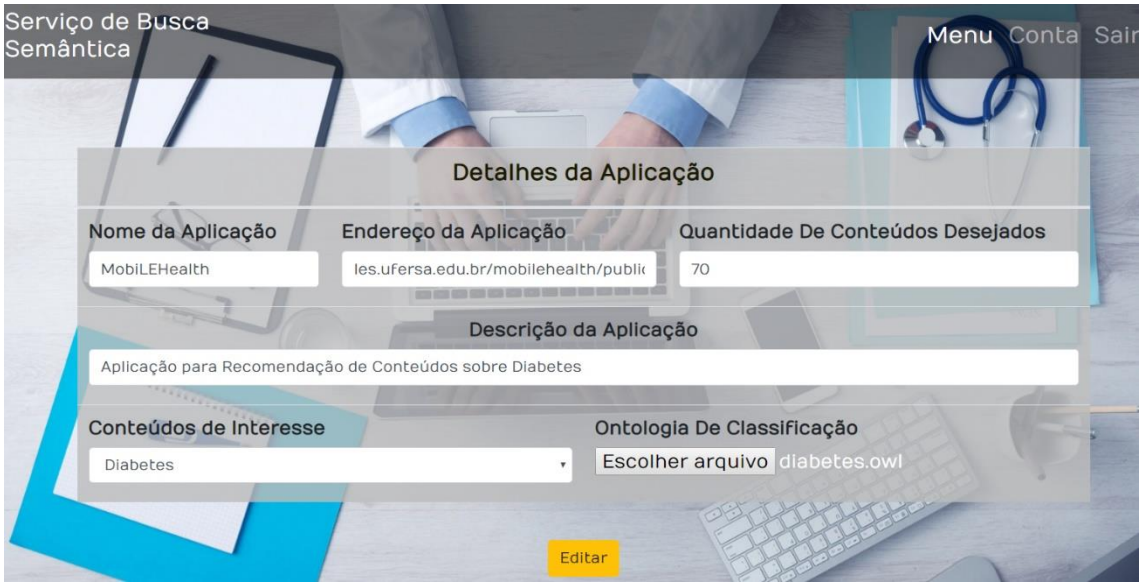
Figura 14 – Tela Principal



Fonte: Autoria Própria

Na tela de cadastro de aplicações, o usuário informa para o motor os dados referentes às aplicações. Nessa etapa, são armazenadas as informações referentes às características da aplicação de saúde, os parâmetros que são necessários para o processo de busca do motor, e a ontologia para classificação dos conteúdos. Esses parâmetros são essenciais para o correto funcionamento do motor. Eles atuam como condições de início e parada do processo de busca. O motor de busca realiza o processo de rastreio e indexação até que as condições sejam satisfeitas. Na figura 15, é apresentada a tela de cadastro de aplicação. Nela, é simulada uma aplicação que requisita o quantitativo de 70 conteúdos e que possuam em seus metadados a palavra DIABETES.

Figura 15 – Tela de Cadastro



Fonte: Autoria Própria

Finalizado o processo de cadastro de aplicação, o gestor da aplicação recebe um e-mail de confirmação, juntamente com os dados necessários para que o mesmo possa inseri-los em sua aplicação de forma possa ter acesso aos conteúdos recuperados pelo motor.

3.3. BUSCA DE CONTEÚDOS

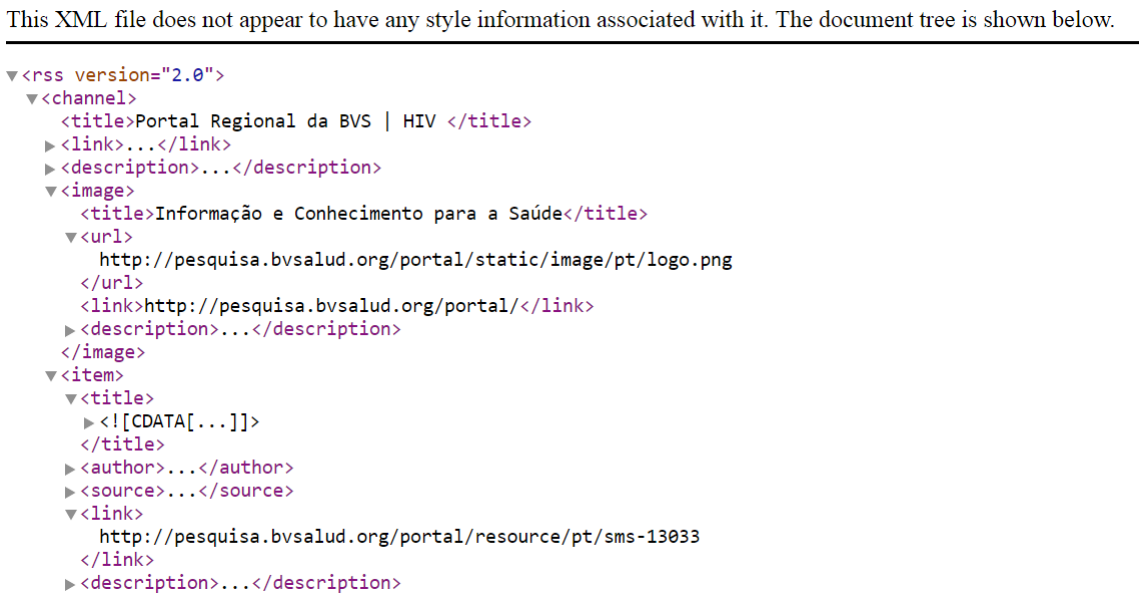
O processo de busca pelos conteúdos é realizado de forma automática e periódica. O motor é configurado para realizar buscas nas bases de dados uma vez por dia com base nas demandas armazenadas em seu banco de dados. Todo o processo de busca pelos conteúdos é realizado no portal da Biblioteca Virtual de Saúde (BVS). A BVS é um site composto por fontes de informação em ciências da saúde que indexa e disponibiliza documentos de bases de dados médicas como MEDLINE, LILACS de forma livre e gratuita. Os conteúdos são disponibilizados em formato de texto, vídeos, manuais de saúde, artigos e teses científicas, abordando todos os tipos de patologias. A seguir, a Figura 16 apresentam o site da BVS juntamente com uma lista de conteúdos.

Figura 16 – Site BVS

Fonte: Autoria Própria

O rastreador realiza o download da página do site da BVS para captura dos conteúdos. O processo de busca é realizado até que os parâmetros de todas as aplicações cadastradas sejam contemplados. Utilizando funções nativas da linguagem PHP, é possível para o motor de busca realizar a leitura do arquivo (site). O site da BVS disponibiliza arquivos XML que contém toda a descrição dos conteúdos armazenados. A seguir, é apresentado na Figura 17, um exemplo de arquivo XML fornecido pelo site. E em seguida, ilustrados os passos do processo de captura de conteúdos.

Figura 17 – Arquivo XML capturado



Fonte: Autoria Própria

- **Verificação:** O rastreador recebe os parâmetros de busca (Quantidade e *String*) da base de dados do motor de busca, e envia uma requisição para o site da BVS com os parâmetros. Esse processo é realizado pela função *file_get_contents* da linguagem PHP. A função realiza o download da página, armazena toda sua estrutura XML em uma variável e inicia a leitura das TAGS;
- **Recepção:** O site da BVS retorna para o rastreador um arquivo .XML com os metadados que descrevem os conteúdos. Dentre as várias tags, são analisadas: Titulo, Descritores, Link de Acesso, Base de Dados Fornecedora;
- **Análise:** O rastreador inicia o processo de leitura das tags do arquivo XML. Através da função *explode* da linguagem PHP, o rastreador procura

pelas tags: <title> que é referente ao título do conteúdo, <url> que é referente ao link para acessar o conteúdo, <db> que é referente a base de dados que fornece o conteúdo, e <mh> que é referente aos metadados que descrevem o conteúdo;

- **Armazenamento:** O rastreador cria um contador temporário que é incrementado à medida que os conteúdos analisados correspondam aos parâmetros de busca. O rastreador verifica dentro da tag <mh> se existem valores correspondentes ao parâmetro de busca (*String*). Caso exista, o rastreador incrementa o contador. Caso não exista, o conteúdo é descartado. O processo de armazenamento é realizado até que o contador tenha o mesmo valor do parâmetro de busca (Quantidade);
- **Envio:** Satisfeito o contador temporário, os conteúdos correspondentes são enviados ao indexador para classificação com base na inferência de conhecimento ontologia;

3.4. INDEXAÇÃO DE CONTEÚDOS

Existem diversas técnicas e abordagens que são capazes de realizar a classificação de documentos pela relação dos mesmos com seus metadados. A abordagem utilizada no motor de busca é baseada em ontologias. O processo de indexação dos conteúdos é realizado pela ontologia. A comunicação do rastreador com a ontologia é realizado pela OWL API. OWL API é uma interface de programação de aplicações de alto nível, desenvolvida em linguagem JAVA, que suporta a criação e manipulação de ontologias. Por meio da API, é possível que o rastreador manipule as ontologias.

Para que a indexação ocorra, é necessário que a ontologia possua alguma métrica para classificação dos indivíduos. Isto é, o rastreador envia os metadados dos conteúdos para a ontologia e recebe como resposta um índice de recomendação. Esse índice deve ser calculado com base na semelhança entre os termos da ontologia e os metadados descritores dos conteúdos. A seguir, é apresentado o fluxo de funcionamento da classificação dos conteúdos para indexação.

- O rastreador envia para o indexador os metadados referentes aos conteúdos analisados;
- O indexador cria indivíduos com características que correspondam aos metadados de cada conteúdo;

- A ontologia realiza leitura das características de cada indivíduo e atribui um índice de recomendação para cada conteúdo;
- Após atribuído índices de recomendação para cada conteúdo, a ontologia insere na base de dados do motor de busca os metadados de cada conteúdo, juntamente com seu nível de recomendação.

3.4.1. ONTOLOGIA PARA CLASSIFICAÇÃO

Com intuito de avaliar e recomendar conteúdos recuperados pelo motor de busca, foi criada uma ontologia que realiza classificação de conteúdos referentes ao diabetes. A descrição da ontologia foi realizada no ambiente Protégé versão 5-4. O Protégé é desenvolvido e mantido pelo Centro de Pesquisa Biomédica da Escola de Medicina da Universidade de Stanford. A linguagem utilizada na especificação da ontologia foi OWL (*Web Ontology Language*). A especificação da ontologia foi orientada pela metodologia *Ontology Development* 101, conforme os passos a seguir.

- DELIMITAÇÃO DO ESCOPO

A ontologia proposta tem por objetivo, classificar conteúdos referentes ao diabetes que são recuperados pelo motor de busca. Para classificação dos conteúdos, a ontologia armazena termos referentes ao diabetes e seus respectivos sinônimos, desta forma, funcionando como um vocabulário controlado. Como forma de representar os conteúdos, a ontologia cria indivíduos que possuem características que representam os metadados que descrevem os conteúdos. A classificação é realizada de acordo com a relação dos metadados com os termos descritos na ontologia. Neste sentido, foram criadas as seguintes questões de competência:

A. Como um conteúdo que possui termos sobre causa, sintoma e tratamento de diabetes pode ser classificado?

B. Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre sedentarismo e obesidade?

C. Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre cicatrização lenta, visão turva e dietas?

D. Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre obesidade, fadiga e insulina?

- REUTILIZAÇÃO DE ONTOLOGIAS

Na literatura foram encontradas algumas ontologias que possuem termos relacionados ao Diabetes. Os autores Lin e Sakamoto (2009) definiram uma ontologia do distúrbio do metabolismo da glicose (OGMD). A ontologia descreve os fatores de suscetibilidade genética relacionados ao diabetes mellitus. Essa ontologia está amplamente relacionada a complicações relacionadas ao diabetes.

Chen et al. (2012) descreveram uma ontologia para medicação do diabetes e uma ontologia para os sintomas dos pacientes. Essa ontologia utiliza a Linguagem de Regras da Web Semântica (SWRL) e Java Expert System Shell (JESS) para determinar possíveis prescrições para pacientes.

Apesar das ontologias apresentarem conceitos referentes ao diabetes, as classes das ontologias não se apresentam como um repositório de sinônimos de maneira que pudessem colaborar com a proposta da presente ontologia. Desta forma, não foi aplicada a etapa de reutilização.

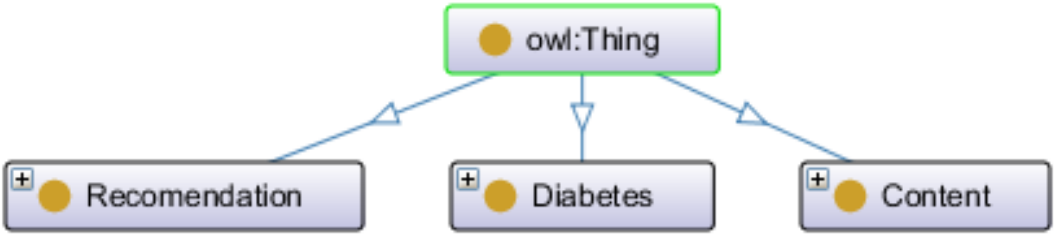
- DEFINIÇÃO DE TERMOS IMPORTANTES

No que diz respeito ao diabetes, os conceitos e termos da presente ontologia são difundidos pela Sociedade Brasileira de Diabetes (SBD) e American Diabetes Association (ADA). Para a presente ontologia, o Diabetes foi definido por causas, sintomas e tratamentos. Como forma de representar a classificação e o nível de recomendação dos conteúdos recuperados foram identificados os termos Conteúdo e Recomendação e Nível de Recomendação.

- DEFINIÇÃO DA HIERARQUIA DE CLASSES

Posteriormente a definição dos termos relevantes, é dado início a criação da hierarquia de classes. As classes foram especificadas utilizando o processo *Top-Bottom.* A presente ontologia possui 25 classes. Dessas, 2 foram criadas com proposito de representar os conteúdos recuperados. 3 classes foram descritas como forma de atribuir um nível de recomendação para cada conteúdo. E por fim, 23 classes foram especificadas com objetivo de representar a síndrome do Diabetes. Na figura 18 são apresentadas as principais classes que são: *Content*, Diabetes e *Recommendation*.

Figura 18 – Principais Classes da Ontologia



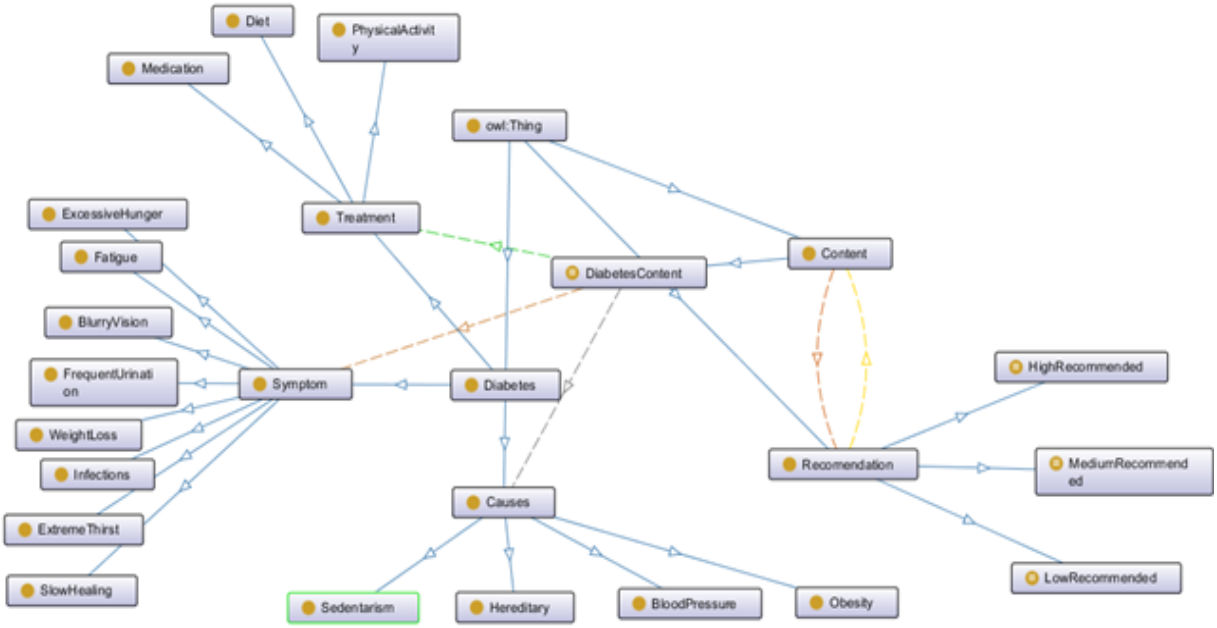
Fonte: Autoria Própria

A classe *Content* foi criada como forma de representar os conteúdos. Essa classe possui uma subclasse (*DiabetesContent)* que indica quais condições um conteúdo precisa satisfazer para ser definido como um conteúdo sobre diabetes.

A classe Diabetes é uma representação dos conceitos referentes ao diabetes. Ela é constituída por 3 subclasses que representam as causas (*Causes*), sintomas (*Symptom*) e tratamentos (*Treatment*). Essa classe é diretamente relacionada com *DiabetesContent*. Para que um conteúdo seja classificado como *DiabetesContent* ele obrigatoriamente precisa possuir relação com alguma subclasse da classe Diabetes.

A classe *Recommendation* foi criada como forma de representar uma partição de valor que atribuirá níveis de recomendação aos conteúdos. Essa classe possui 3 subclasses (*LowRecommended*, *MediumRecommended* e *HighRecommended*) que retratam os possíveis níveis de recomendação. Os níveis são atribuídos conforme a relação dos metadados com as subclasses da classe Diabetes. Na Figura 19, é apresentada a hierarquia de classes da ontologia.

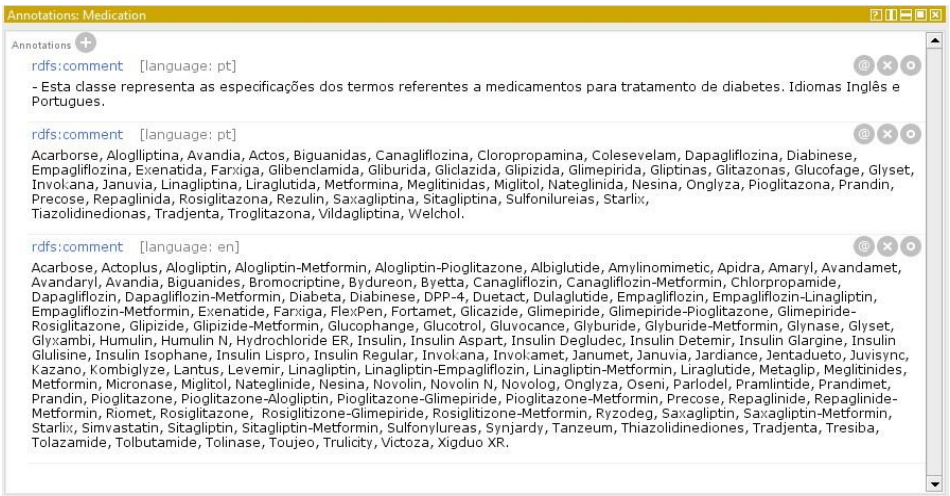
Figura 19 – Hierarquia de Classes



Fonte: Autoria Própria

É importante destacar que todas as subclasses de Diabetes possuem um repositório de sinônimos. O objetivo desse repositório é enriquecer os termos de maneira que, no momento da consulta à ontologia, seja considerada toda a descrição da classe e não somente o nome. Na Figura 20, é ilustrada a descrição da classe medicamento (*Medication*).

Figura 20 – Descrição da Classe Medication



Fonte: Autoria Própria

Primeiramente, a ontologia verifica se os metadados são descritos com algum termo referente ao diabetes. Caso haja, o conteúdo é classificado como conteúdo de diabetes (*DiabetesContent*). Caso não haja, é classificado como conteúdo (*Content*) e logo descartado. Em seguida é verificada a recorrência em que as subclasses de Diabetes são representadas no conteúdo e calculado o nível de recomendação.

A ontologia considera que, caso haja representação das três subclasses (*Causes*, *Symptom* e *Treatment*), o conteúdo é classificado como muito recomendado (*HighRecommended*). No cenário onde sejam representadas duas subclasses distintas, por exemplo: Treatment e *Symptom*, a classificação é apresentada como parcialmente recomendada (*MediumRecommended*). Pouco recomendada (*LowRecommended*) é a classificação referente à aparição de somente uma representação de alguma das subclasses.

- DEFINIÇÃO DAS PROPRIEDADES

A criação dos relacionamentos entre classes e indivíduos foi realizada por meio das propriedades de objeto (*Object Property*). As propriedades trabalham com o conceito de domínio e imagem. Essas propriedades representam relacionamentos entre duas classes ou dois indivíduos, unindo indivíduos ou classes de um domínio a indivíduos ou classes de uma imagem. Na Tabela 5, são apresentadas as propriedades, suas descrições e seus respectivos domínios e imagens.

Tabela 5 – Descrição das Propriedades

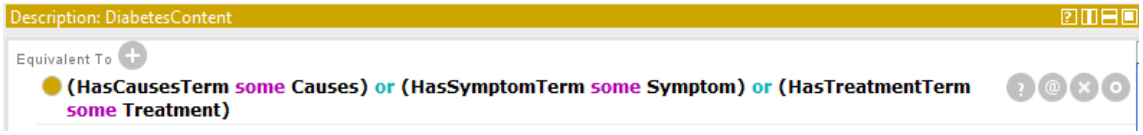
Propriedade	**Descrição**	**Domínio**	**Imagem**
Has_CausesTerm	Essa propriedade indica que o conteúdo tem algum termo referente a causa de diabetes.	*DiabetesContent*	*Causes*
Has_SymptomTerm	Essa propriedade indica que o conteúdo tem algum termo referente a sintoma de diabetes.	*DiabetesContent*	*Symptom*
Has_TreatmentTerm	Essa propriedade indica que o conteúdo tem algum termo referente a tratamentos de diabetes.	*DiabetesContent*	*Treatment*

Has_Recommendation	Essa propriedade indica que o conteúdo possui algum nível de recomendação.	*DiabetesContent*	*Recommendation*

- DEFINIÇÃO DAS CARACTERÍSTICAS

As definições de características permitem criar condições para as classes. Para que um indivíduo seja considerado membro de determinada classe, ele precisa satisfazer essas condições. Na figura 21, é apresentada a classe *DiabetesContent* e sua condição de restrição. Para que um indivíduo seja classificado como um conteúdo de diabetes (*DiabetesContent*) ele precisa ter pelo menos uma relação com as subclasses causa (*Causes*) por meio da propriedade *HasCausesTerm*, ou sintomas (*Symptom*) por meio da propriedade *HasSymptomTerm*, ou tratamento (*Treatment*) por meio da propriedade *HasTreatmentTerm*.

Figura 21 – Classe DiabetesContent



Fonte: Autoria Própria

Para que um indivíduo seja classificado como muito recomendado (*HighRecommended*) ele precisa ser obrigatoriamente um conteúdo de diabetes (*DiabetesContent*) e possuir relação com as 3 Classes: causa (*Causes*), sintoma (*Symptom*) e tratamento (*Treatment*), por meio de suas respectivas propriedades. Na figura 22, é apresentada a classe *HighRecommended* e sua condição de restrição.

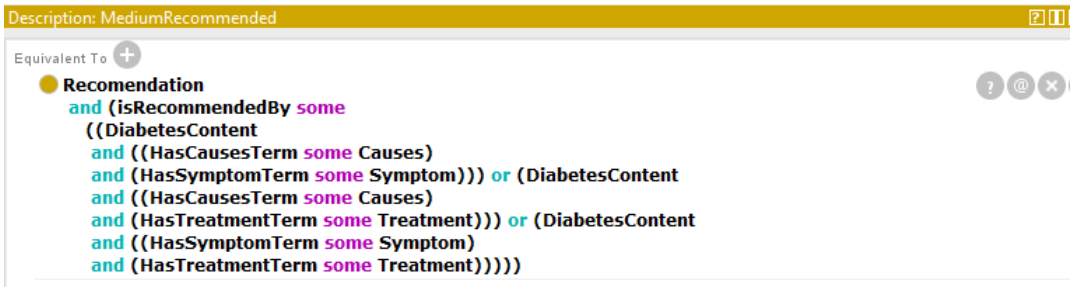
Figura 22 – Classe HighRecommended

Fonte: Autoria Própria

Para que um indivíduo seja classificado como parcialmente recomendado (*MediumRecommended*) ele precisa ser obrigatoriamente um conteúdo de diabetes (*DiabetesContent*), e possuir pelo menos uma relação com exatamente 2 classes distintas, por meio de suas respectivas propriedades. Por exemplo: um conteúdo que tenha termos referentes a sintomas (*Symptom*) e tratamento (*Treatment*) será classificado como parcialmente recomendado (*MediumRecommended*).

É importante salientar, caso, o conteúdo possua mais de uma relação com a mesma subclasse, é contabilizado somente uma. Por exemplo: um conteúdo que tenha 4 relações com causas (*Causes*) e 5 com sintomas (*Symptom*), será classificado como parcialmente recomendado (*MediumRecommended*). Na figura 23, é ilustrada a classe *MediumRecommended* e suas restrições.

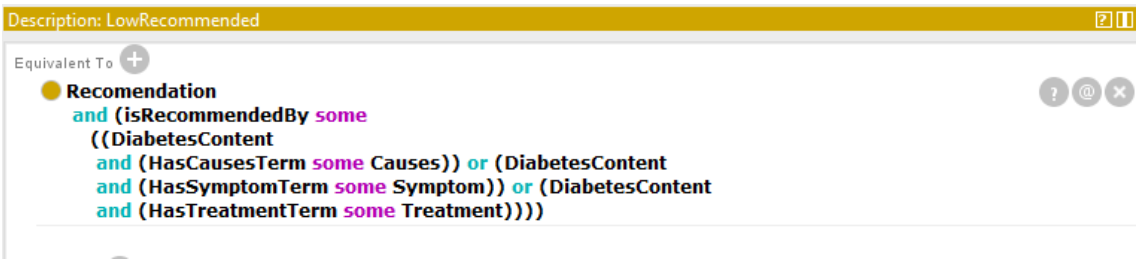
Figura 23 – Classe MediumRecommended



Fonte: Autoria Própria

Para que um indivíduo seja classificado como pouco recomendado (*LowRecommended*) ele precisa ser obrigatoriamente um conteúdo de diabetes (*DiabetesContent*) e possuir relação com somente uma única subclasse, por meio de suas respectivas propriedades. Por exemplo: um conteúdo que tenha termos referentes a tratamento (*Treatment*) será classificado como pouco recomendado (*LowRecommended*). Na figura 24, é apresentada a classe *LowRecommended* e suas restrições.

Figura 24 – Classe LowRecommended



Fonte: Autoria Própria

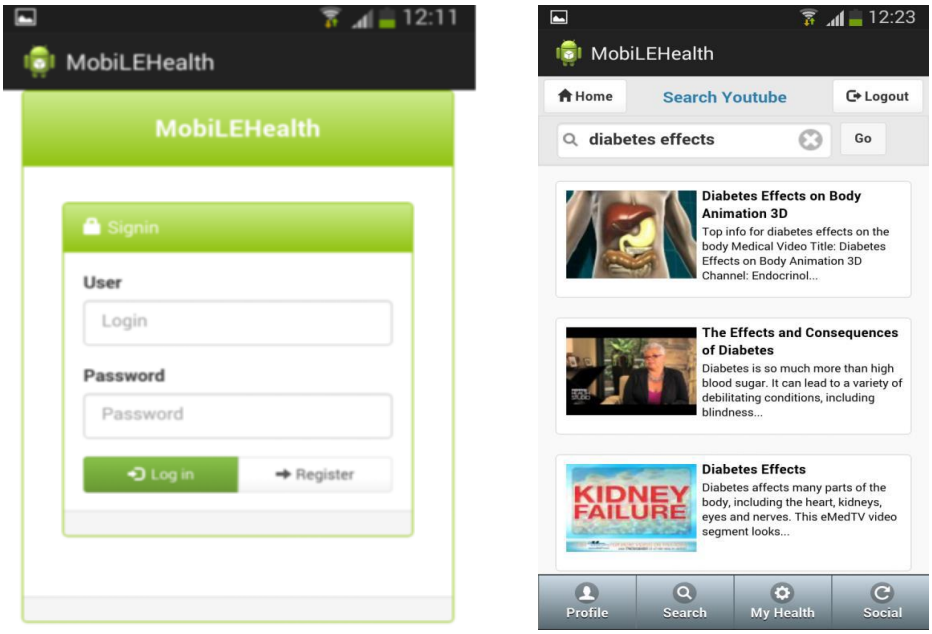
- ELABORAÇÃO DE INSTÂNCIAS

Para o correto funcionamento da presente ontologia, não há a necessidade de existência de indivíduos criados previamente. Os indivíduos são criados de acordo com as requisições enviadas pelo motor de busca. De forma resumida, o motor envia para ontologia os metadados referentes a um conteúdo, a ontologia cria os indivíduos conforme a descrição dos mesmos, realiza a classificação e retorna o nível de recomendação.

3.5. INTEGRAÇÃO

Para execução de testes relacionados à captura de conteúdos, foi realizada integração do motor de busca com a aplicação MobiLEHealth. O ambiente *Mobile Learning Environment for Health*, denominado "MobiLEHealth", consiste em um ambiente de aprendizagem ubíqua e informal no contexto de Saúde 2.0 destinado a pessoas com doenças crônicas. O MobiLEHealth pode ser acessado pelo endereço: http://les.ufersa.edu.br/mobilehealth/public/users/login. A figura 25 exibe a tela inicial e a tela de acesso a vídeos do youtube na aplicação MobileHealth.

Figura 25 – Telas MobiLEHealth



Fonte: Autoria Própria

Como descrito na seção 3.2, ao final do cadastro da aplicação, o gestor da aplicação recebe os dados necessário para que possa receber os conteúdos recuperados pelo motor de busca em sua aplicação. Na Figura 26, mais especificamente na linha 5, são apresentados os dados necessários para conexão.

Figura 26 – Conexão com o Motor

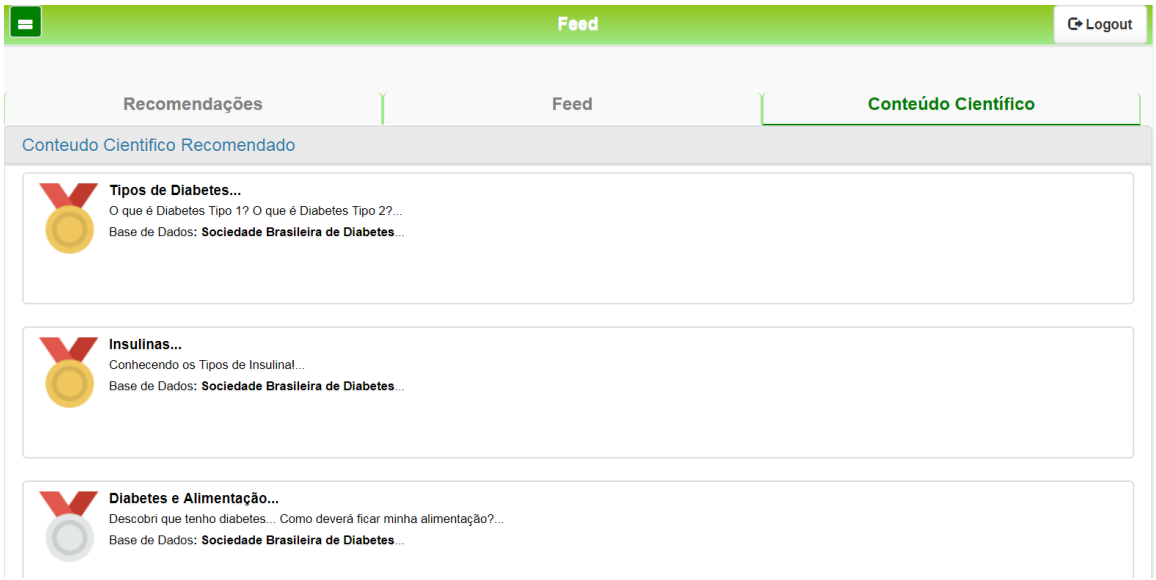
```php
<?php

    $recebe = $_SERVER["REQUEST_URI"];
    $idcontent = substr($recebe, 34, strlen($recebe));
    $conexao = mysqli_connect('localhost', 'root', 'z800lesi514', 'sss');
    $sql = mysqli_query($conexao, "select * from content where idcontent = $idcontent;");
    $registro = mysqli_fetch_array($sql);
    $url = $registro['link'];
?>
```

Fonte: Autoria Própria

Além dos dados para conexão com o motor, para que pudesse receber os conteúdos capturados, foi necessário realizar algumas alterações no MobiLEHealth. Foi criada uma tela denominada conteúdos científicos que exibe os conteúdos capturados pelo motor de busca. Na figura 27, é ilustrado os dados necessários para realizar a conexão entre o motor de busca e o mobiLEHealth.

Figura 27 – Tela de Conteúdos Científicos



Fonte: Autoria Própria

As medalhas posicionadas à esquerda do título de cada conteúdo representam os níveis de recomendação gerados pela ontologia. A ontologia será descrita na próxima seção desse capitulo. Os conteúdos possuem 3 níveis de recomendação. Conteúdos muito

recomendados são representados pelas medalhas de ouro. Os parcialmente recomendados são ilustrados pela medalha de prata. E, por fim, os pouco recomendados são representados pelas medalhas de bronze. Na Figura 28, é exibido o acesso a um dos conteúdos disponibilizados pelo motor.

Figura 28 – Conteúdo Recomendado

Fonte: Autoria Própria

4. EXPERIMENTOS E VALIDAÇÕES

O presente capítulo apresenta validações para os resultados obtidos pela prototipação. O objetivo dos testes, é medir e avaliar a capacidade do motor de busca de recuperar os conteúdos armazenados na base de dados BVS. Dentro os diversos assuntos contidos nos conteúdos, foram considerados os documentos que eram descritos pelos metadados como conteúdos de diabetes. Inicialmente são apresentados os cálculos das métricas de recuperação de informação a respeito dos conteúdos capturados. Em seguida são ilustrados os passos da avaliação da ontologia para classificar os conteúdos recuperados. E por fim, são registrados os resultados de um questionário a respeito da qualidade dos conteúdos disponibilizados.

4.2. TESTES EXPERIMENTAIS

Como forma de analisar a qualidade do processo de busca e recuperação das informações do motor, foram utilizadas as métricas *Recall* e *Precision*. Com essas duas métricas, é possível mensurar o número de documentos relevantes que foram recuperados (*Recall*). Além disso, é possível avaliar, dentre os documentos recuperados, a quantidade dos documentos que são realmente relevantes para busca (*Precision*).

Foi utilizada uma amostra de 80 conteúdos. Foram avaliados os resultados da execução do motor de busca em 80 documentos descritos como conteúdos de diabetes, pela BVS. O do objetivo do teste foi medir a eficácia da busca. Os resultados foram avaliados utilizando as métricas de revocação e precisão. A revocação foi utilizada para medir a habilidade do sistema em recuperar os documentos mais relevantes para o usuário, ou seja, mede-se o coeficiente entre a quantidade de itens relevantes que foram recuperados e total de itens relevantes existentes na base de dados. Após execução da busca, o motor identificou que todos os 80 conteúdos da amostra eram descritos como conteúdos de diabetes. Aplicando a métrica precisão, temos:

$$Revocação = \frac{Documentos\ Relevantes\ Recuperados}{Total\ de\ Documentos\ Relevantes} \times 100$$

$$Revocação = \frac{80}{80} \times 100 = 100\%$$

Visando complementar o processo de validação das buscas, foi utilizada a métrica de avaliação de sistemas de busca. A precisão, por sua vez, mede a habilidade do sistema de manter os documentos irrelevantes fora do resultado de uma consulta, ou seja, mede-se a quantidade de itens relevantes dentre os itens retornados para a consulta. Para utilização da métrica precisão, é necessário atribuir um ou mais parâmetros de relevância, para analisar se os documentos recuperados, são correspondentes à métrica. Dessa forma, foi definido que para que um conteúdo seja relevante para a busca, ele precisa ser descrito em idioma português. Assim, foi utilizado o mesmo intervalo de 80 conteúdos. Dentro do intervalo estabelecido, o motor realizou a leitura dos metadados de cada conteúdo, e no comparativo com o termo “Diabetes” e que eram descritos em idioma português, identificou que 62 eram atendiam aos parâmetros. Aplicando a métrica precisão temos:

$$Precisão = \frac{Total\ de\ Documentos\ Relevantes\ Recuperados}{Total\ de\ Documentos\ Recuperados} \times 100$$

$$Precisão = \frac{62}{80} \times 100 = 77{,}5\%$$

Os resultados obtidos foram considerados satisfatórios. Com base na métrica recall, o motor foi capaz de interpretar que todos conteúdos eram referentes ao contexto do diabetes. A métrica precisão, também foi considerado satisfatório para o processo de captura dos conteúdos, visto que mais de dois terços do intervalo foram identificados como relevantes.

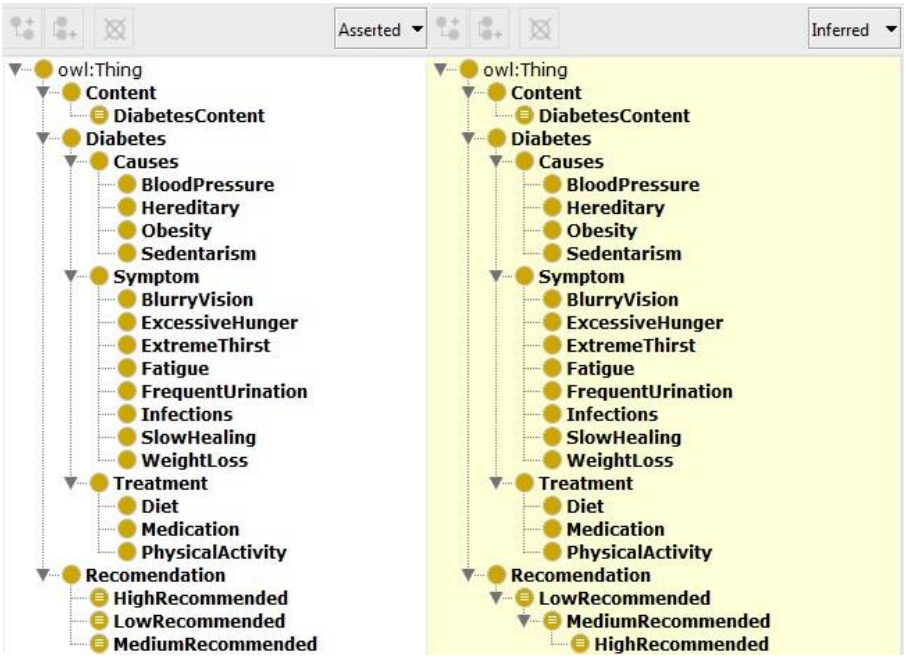
4.3. AVALIAÇÃO DA ONTOLOGIA

Como qualquer artefato tecnológico, ontologias precisam ser avaliadas. O propósito da avaliação é estimar a qualidade da ontologia. O autor Gómez-Pérez (2004) sugere duas abordagens para avaliação de ontologias: verificação e validação.

4.3.1. VERIFICAÇÃO

O processo de verificação é dividido em duas etapas. Na primeira, procura-se analisar a consistência das classes a fim de verificar se as definições da ontologia estão concisas. Para responder este critério, foram utilizados os *reasoners* da ferramenta Protégé. Caso exista inconsistência em alguma classe, após a inferência, o *reasoner* destaca em cor vermelha a classe inconsistente. Na Figura 29, são exibidos dois cenários. No primeiro, a hierarquia de classes afirmativas (*Asserted*). No segundo, o resultado da inferência, classes inferidas (*Inferred*). É possível visualizar que não existe inconsistência nas classes.

Figura 29 – Visualização das Classes Afirmativas/Inferidas



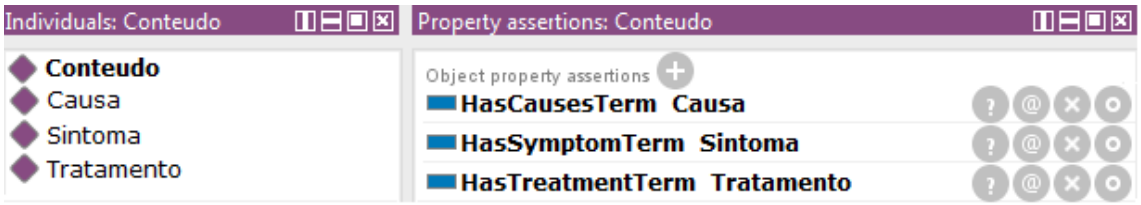
Fonte: Autoria Própria

A segunda etapa da verificação é realizada por meio das repostas para as questões de competência (QC) previamente elaboradas. Caso a ontologia seja capaz de responder corretamente as QC, é possível verificar que suas definições e relações estão coerentes com o domínio especificado. Para essa etapa, foram criados indivíduos hipotéticos que simulam conteúdos recuperados e metadados que os descrevem. A seguir, serão apresentadas as questões de competências e suas respectivas respostas.

QC 1 - Como um conteúdo que possui termos sobre causa, sintoma e tratamento de diabetes pode ser classificado?

A questão de competência 1 aborda a relação entre os conteúdos e sua descrição. Na Figura 30, é realizada uma simulação com 4 indivíduos que representam um conteúdo recuperado e 3 metadados que o descrevem. O indivíduo Conteúdo se relaciona com o indivíduo Causa por meio da propriedade *HasCausesTerm*. Da mesma forma, se relaciona com o indivíduo sintoma através da propriedade *HasSymptomTerm*. E, por fim, se relaciona com o indivíduo tratamento pela propriedade *HasTreatmentTerm*.

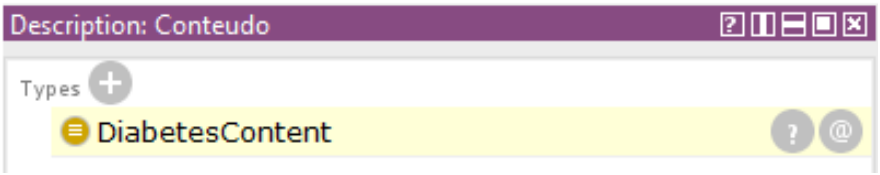
Figura 30 – Primeira Simulação



Fonte: Autoria Própria

Após execução do motor de inferência, a ontologia classificou que o indivíduo Conteúdo pertence a classe *DiabetesContent*, respondendo assim a primeira QC de maneira correta. Na figura 31, é ilustrado o resultado da inferência.

Figura 31 – Resultado da Primeira Simulação

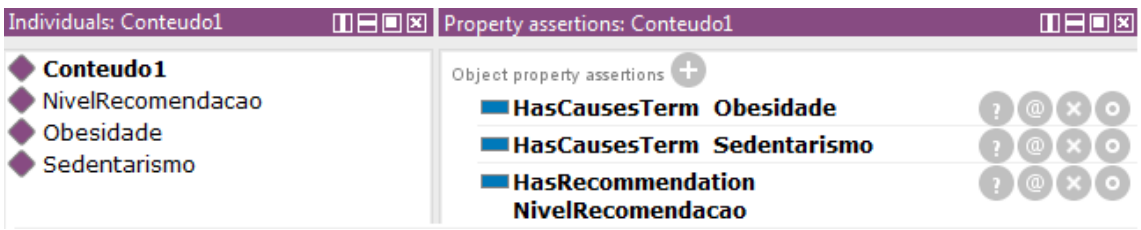


Fonte: Autoria Própria

QC 2 - Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre sedentarismo e obesidade?

As questões 2, 3 e 4 foram criadas com intuito de verificar se a ontologia é capaz de atribuir níveis de recomendação, de acordo com a recorrência em que as subclasses da classe diabetes são representadas nos conteúdos. Na Figura 32, é realizada uma simulação com 4 indivíduos. O indivíduo Conteúdo1 representa um conteúdo recuperado. O indivíduo NivelRecomendação representa o nível que será atribuído ao conteúdo após inferência. O outros dois, representam metadados que descrevem o conteúdo.

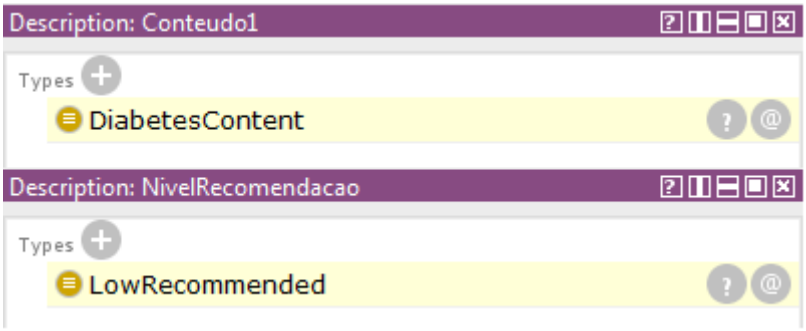
Figura 32 – Simulação 2



Fonte: Autoria Própria

Durante a inferência, a ontologia classificou que o indivíduo Conteudo1, é membro da classe *DiabetesContent* e possui um nível pouco recomendado (*LowRecommended*). Apesar de possuir 2 relações, ambas são com a mesma subclasse (*Causes*). Na figura 33, é apresentado o resultado da inferência.

Figura 33 – Resultado Simulação 2



Fonte: Autoria Própria

QC 3 - Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre cicatrização lenta, visão turva e dietas?

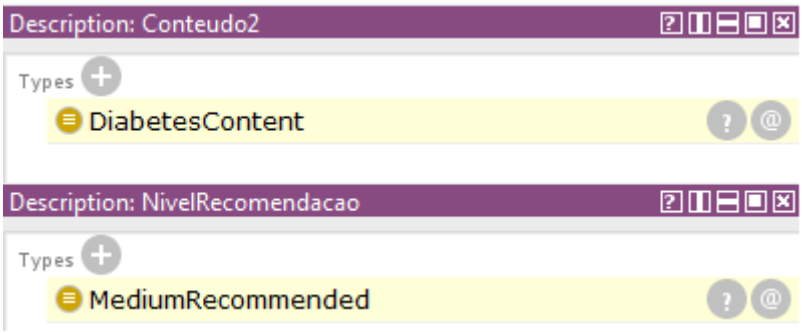
Na terceira simulação, foram criados 5 indivíduos. O indivíduo Conteudo2, por meio das propriedades *HasSymptom* e *HasTreatment*, se relaciona com outros indivíduos (Dietas e VisaoTurva), e, através da propriedade *HasRecommendation*, se associa a um nível de recomendação (NivelRecomendacao). A Figura 34 apresenta os indivíduos e suas relações.

Figura 34 – Simulação 3

Fonte: Autoria Própria

Como resultado da inferência tem-se que o indivíduo Conteudo2 é membro da classe *DiabetesContent*. Seu nível de recomendação foi classificado como parcialmente recomendado (*MediumRecommended*), devido às suas relações com 2 subclasses distintas. Na Figura 35, é demonstrado o resultado da inferência.

Figura 35 – Resultado Simulação 3

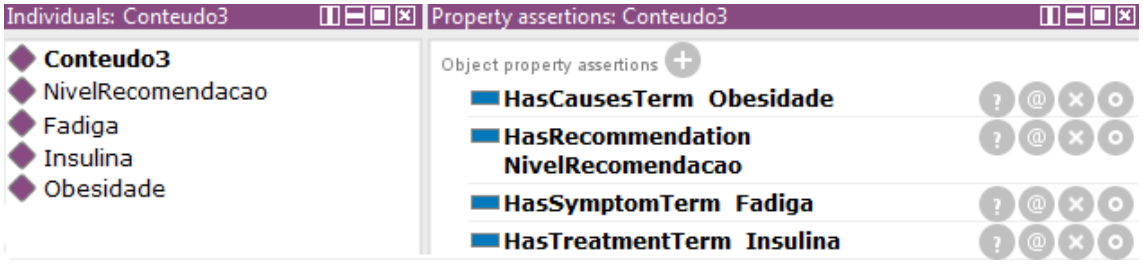


Fonte: Autoria Própria

QC 4 - Qual nível de recomendação de um conteúdo de diabetes que possui termos sobre fadiga, insulina e obesidade?

A última simulação ilustra um indivíduo (Conteudo3) que possui relação com todas as 3 subclasses principais da classe Diabetes. Mediante as propriedades *HasCauses*, *HasSymptom* e *HasTreatment*, o conteúdo se relaciona com os outros indivíduos. Na Figura 36, são descritos os indivíduos e relações.

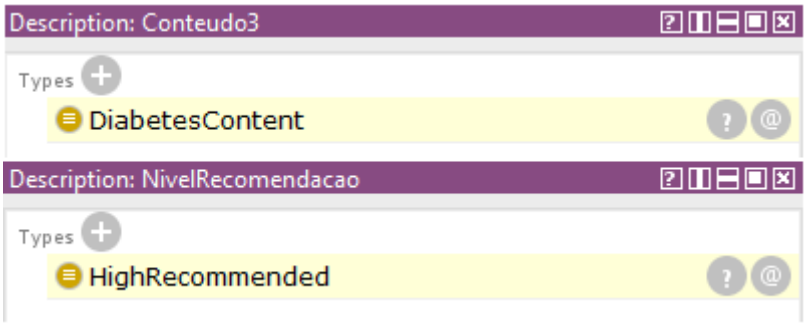
Figura 36 – Simulação 4



Fonte: Autoria Própria

A seguir, na Figura 37, é apresentado o resultado da inferência. Devido às suas relações com todas as subclasses de Diabetes, o indivíduo Conteudo3 foi classificado como *DiabetesContent* e possui um nível muito recomendado (*HighRecommended*).

Figura 37 – Resultado Simulação 4



Fonte: Autoria Própria

A formulação das questões de competência permitiu traduzir, de forma hipotética, situações que representem casos relacionados à realidade de indivíduos diabéticos. As questões apresentaram-se de maneira concisas em relação ao contexto de Diabetes, visto que os termos presentes em cada questão estão diretamente ligados ao domínio do diabetes. As inferências realizadas pela ontologia se mostraram coerentes para responder as questões de competência previamente estabelecias.

4.3.2. VALIDAÇÃO

A fim de avaliar as classificações realizadas pela ontologia, foram conduzidos experimentos com conteúdos disponibilizados pelo motor de busca. A ontologia proposta foi integrada ao motor com o intuito de classificar os conteúdos de acordo com a relação dos metadados descritores com Domínio de Diabetes. Foram analisados 10 conteúdos recuperados pelo motor. O objetivo do experimento foi apresentar uma validação prática utilizando casos reais. Por meio da análise dos resultados, procurou-se avaliar a precisão da classificação. A Tabela 6 apresenta alguns conteúdos retornados pelo ambiente de aprendizagem e suas respectivas classificações.

Tabela 6 – Conteúdos Recuperados e Classificados

Título Conteúdo	Metadados	Classificação
Diabetes: causa, sintomas, tratamentos	Diabetes, Causas, Sintomas, Tratamentos, Endocrinologia, Metabólica, CRM	*HighRecommended*
O que é Diabetes?	Classificação, Diabetes, Medicina, Insulina, Gestação	*MediumRecommended*
Diabetes - O que é e como tratar	Informações, Diabetes, Prevenção, Conteúdos	*LowRecommended*
Tudo sobre Sintomas da Diabetes: Como Tratar, Controlar e Reverter	Organismo, Patologia, Glicemia, Alimentação	*MediumRecommended*
Diabetes: Não Ignore Estes Primeiros Sintomas	Sintomas, Hipertensão, Aferimento, Médicos, Especialistas	*LowRecommended*
Curso de Farmacologia: Aula 25 - Diabetes - Fisiopatologia	Fisiopatologia, Imunologia, Especialistas, Diabetes	*LowRecommended*
7 Sintomas da diabetes – Conheça os principais Sintomas da Diabetes	Diabetes, Sintomas, Tratamentos, Causas, Aprendizagem	*HighRecommended*
Diabetes CONTROLADA - Acabe com a Diabetes de Forma 100% Natural e Glicose Controlada	Diabetes, Controle, Mellitus, Tipos, Homeopatia	*MediumRecommended*
Notícias sobre Diabetes EXAME	Revista, Publicação, Diabetes, Conhecendo	*LowRecommended*
Diabetes - Como cai na Prova? - Concursos de Enfermagem	Aprendizagem, Diabetes, Conhecimento, Enfermagem	*LowRecommended*

Perante os resultados apresentados, foi possível avaliar a precisão da classificação dos conteúdos. Durante os experimentos foi possível notar que o nível de recomendação está diretamente ligado a quão bem são descritos os metadados. Também foi possível verificar que a quantidade de sinônimos armazenados nas classes influencia diretamente na classificação, pois uma maior quantidade de sinônimos, consequentemente, eleva a probabilidade de uma melhor classificação.

4.4. QUESTIONÁRIO DE AVALIAÇÃO DE CONTEÚDOS

O questionário tem por objetivo medir a satisfação dos usuários do sistema com relação aos conteúdos recomendados. Para criação do questionário foi utilizada a ferramenta de criação de formulários Google Forms. Foram elaboradas 10 perguntas que possuem relação direta com os conteúdos disponibilizados. Em 9 perguntas, foi utilizada

a escala Likert para medir a concordância da resposta do usuário com a pergunta. A escala Likert oferece as seguintes opções: Discordo Totalmente, Discordo Parcialmente, Não Concordo, Nem Discordo, Concordo Parcialmente e Concordo Totalmente.

Responderam o questionário, um total de 12 pesquisadores da área da saúde com ênfase no diabetes. O experimento foi conduzido em um intervalo de 13 dias. A faixa etária dos entrevistados concentrou-se entre 18 e 24 anos, como pode ser observado no gráfico 1.

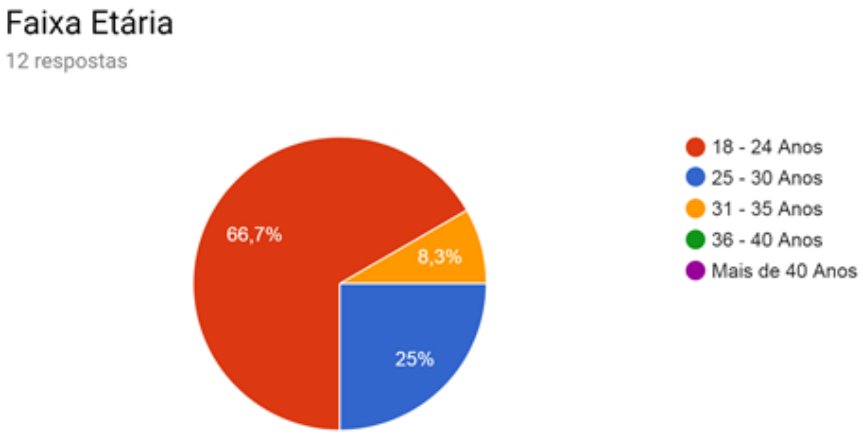


Gráfico 1 – Faixa Etária dos Pesquisadores

Como forma de acompanhar a frequência de acessos dos usuários, foi criada a pergunta que indaga quantos conteúdos o usuário leu. Conforme o gráfico 2, é possível identificar uma amostragem variada, visto que todas opções foram contempladas. No gráfico 2, são apresentadas as respostas para a presente pergunta.

Gráfico 2 – Leitura dos Conteúdos

No gráfico 3, são apresentadas as respostas dos usuários quanto à percepção dos mesmos em relação à recorrência de termos referentes ao Diabetes nos conteúdos. É possível identificar que mais de 80% das respostas concordam que os conteúdos apresentam termos sobre diabetes.

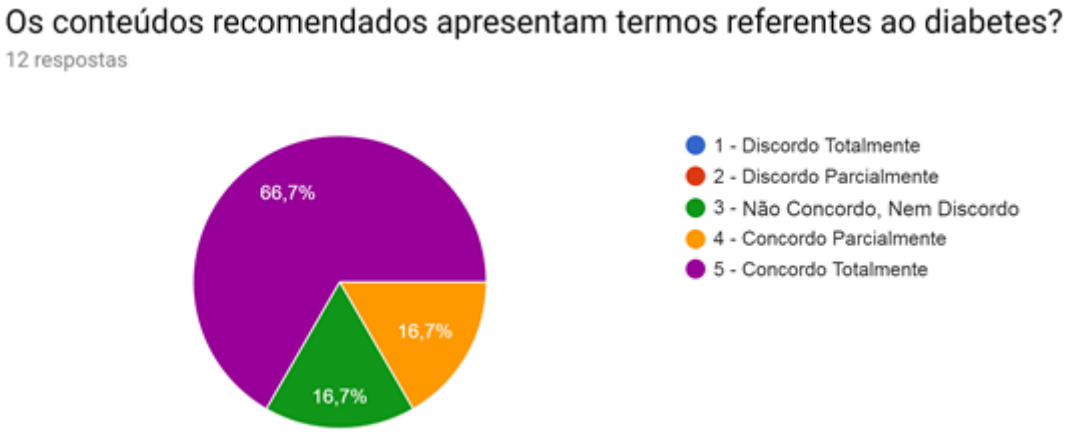


Gráfico 3 – Termos Referentes ao Diabetes

A intenção na elaboração da próxima pergunta foi identificar a opinião dos especialistas, quanto à leitura dos conteúdos. Por meio das respostas, mais da metade dos usuários concordaram que os conteúdos se apresentaram de alguma forma como interessantes. No gráfico 4, são ilustradas as respostas.

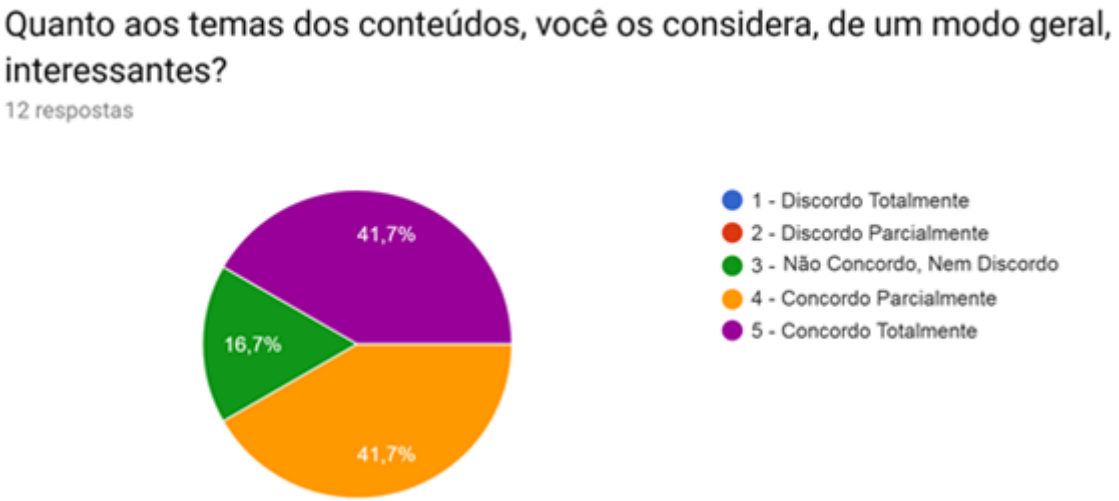


Gráfico 4 – Assuntos Interessantes Abordados nos Conteúdos

A pergunta seguinte procura comparar as respostas dos usuários com o nível de recomendação que é atribuído aos conteúdos pela ontologia. Foi possível identificar que metade dos pesquisadores concorda com o nível de recomendação atribuído aos conteúdos. A outra metade não concordou e nem discordou. No gráfico 5, são apresentadas as respostas.

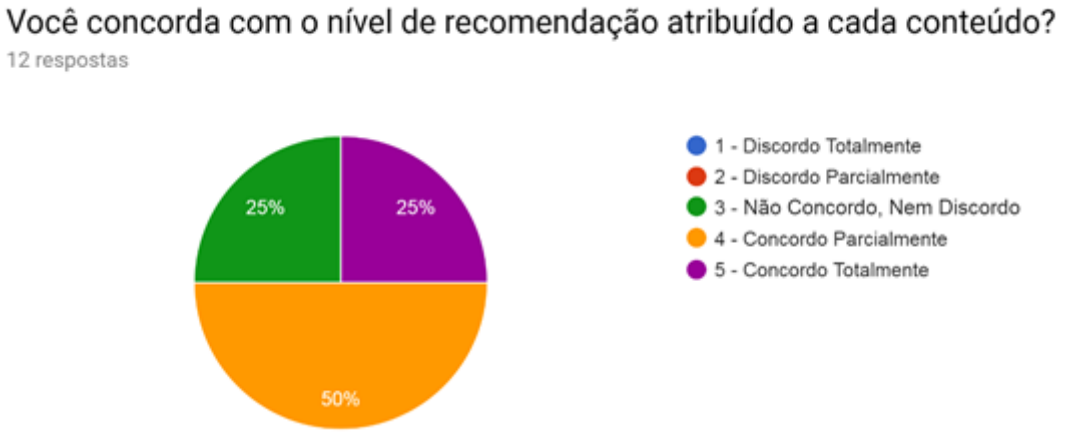


Gráfico 5 – Nível de Recomendação dos Conteúdos

Quanto à possibilidade de os conteúdos contribuírem de alguma forma para o tratamento de diabetes, 80% dos usuários responderam que acreditam que sim. Os outros 16,7% não concordaram e nem discordaram. No gráfico 6, são ilustradas as respostas dos pesquisadores.

Gráfico 6 – Auxílio no Tratamento de Diabetes

Conforme apresentado no Gráfico 7, a maioria dos usuários concordam que os conteúdos permitem aos usuários mais clareza quanto ao tema. Cerca de 83,3% dos entrevistados concordam parcialmente ou totalmente com essa afirmação.

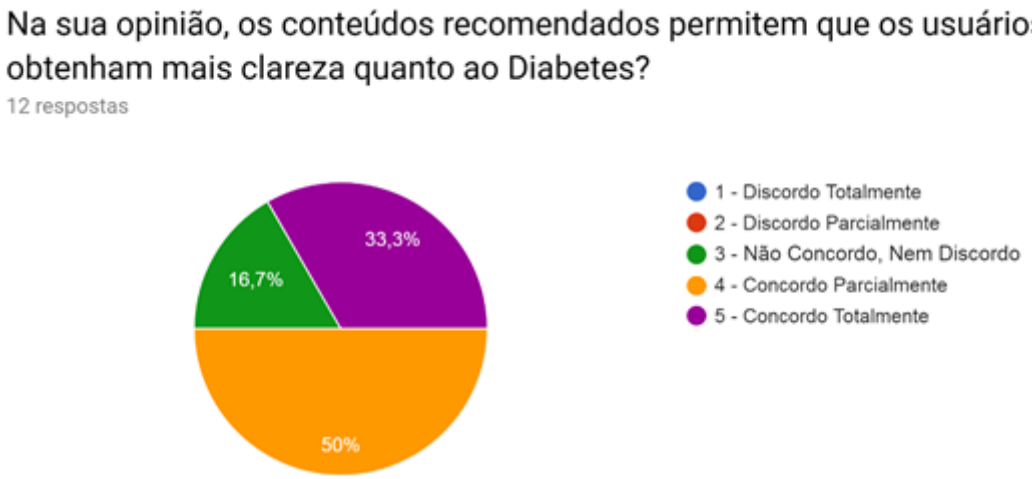


Gráfico 7 – Clareza Quanto ao Tema

Como forma de verificar se os conteúdos disponibilizados estão de acordo com o tema, foi perguntado aos entrevistados se eles concordavam que os conteúdos eram adequados ao diabetes. O gráfico 8 ilustra que 75% concordam que os conteúdos estão de acordo com o tema.

Gráfico 8 – Conteúdos Adequados Quanto ao Tema

No gráfico 9, foi questionado aos entrevistados se a quantidade de conteúdos que foram disponibilizados era satisfatória. De acordo com as respostas, 75% dos

pesquisadores concordam que o número de conteúdos é satisfatório. No gráfico 9, são apresentadas as respostas.

Gráfico 9 – Satisfação Quanto à Quantidade de Conteúdos

Um dos princípios da disponibilização de informações de saúde na web, é o de que a população possa ter maior conhecimento sobre seu atual estado de saúde e, consequentemente, uma melhor qualidade de vida. Pensando nisso, foi elaborada a próxima pergunta. Assim, mais de 75% das respostas concordam que os conteúdos contribuem para uma melhor qualidade de vida.

Gráfico 10 – Melhor Qualidade de Vida

E, por fim, foi perguntado aos entrevistados, se os mesmo indicariam os conteúdos disponibilizados para pacientes e profissionais da saúde. Segundo as respostas, mais de 80% dos entrevistados indicariam os conteúdos.

Gráfico 11 – Indicação de Conteúdos

Com base nas respostas dos pesquisadores, é possível visualizar um cenário positivo quanto aos conteúdos recomendados. Por se tratar da opinião de especialistas na área do diabetes, os resultados se apresentam de maneira satisfatória e têm-se que o motor cumpre seu papel de fornecer conteúdos de confiáveis para aplicações de saúde.

5. CONSIDERAÇÕES FINAIS

O presente trabalho apresentou um motor de busca para dados de saúde na web. Por intermédio do motor, é possível alimentar aplicações de saúde (*e-health*), com informações advindas de bases de dados médicas. O motor responde as requisições das aplicações de saúde, disponibilizando conteúdos que correspondam aos parâmetros passados. Como forma de utilização prática em ambiente real, foi realizada a integração do motor com o sistema de recomendação MobiLEHealth. O motor recebeu os parâmetros passados e retornou parou o MobiLEHealth uma lista de conteúdos referentes ao diabetes.

Para filtragem e disponibilização de conteúdos que correspondessem aos parâmetros passados pelo MobiLEHealth, foi especificada e desenvolvida uma ontologia, para classificação de conteúdos voltados ao contexto do diabetes. A utilização da Metodologia 101 possibilitou a criação de uma ontologia que respondesse às questões de pesquisa. Os passos iterativos da metodologia permitiram que, em cada fase da especificação, a ontologia fosse refinada até uma versão satisfatória. A abordagem utilizando a ontologia como um vocabulário controlado permitiu enriquecer as classes

com seus respectivos sinônimos. A ontologia utilizada demonstrou ser uma forte aliada na classificação dos dados.

As bases de dados de saúde encontradas apresentaram-se como fontes de informações confiáveis. O fato de serem mantidas por profissionais de saúde eleva o índice de credibilidade dos conteúdos que chegam até os usuários. Além de serem repositórios de grande conhecimento não somente em relação ao diabetes como qualquer ramificação no domínio da saúde.

Quanto à aplicação do questionário, a avaliação dos pesquisadores da área de saúde foi positiva. De forma geral, as respostas dos entrevistados sobre a qualidade e relevância dos conteúdos foi bastante satisfatória, indicando um caminho correto para prosseguimento futuro do trabalho.

5.1. LIMITAÇÕES

A principal limitação do motor se caracteriza pelo aumento gradativo do esforço computacional, à medida que cresce a o número de conteúdos capturados. Caso o motor receba um parâmetro alto para a quantidade de conteúdos, o processo de rastreamento precisa realizar muitas requisições ao portal da BVS.

Outra limitação é em relação à forma como são capturados os conteúdos. Por mais que os metadados sejam especificados por profissionais da área de saúde, uma descrição desses metadados realizada de forma errada acarreta em uma não-captura do conteúdo ou, até mesmo, a disponibilização de conteúdos que não correspondam de fato à busca.

Em relação à ontologia, a principal limitação identificada fica por conta da não-lematização dos termos de forma automática. A necessidade de descrever manualmente cada classe e suas respectivas variações demanda muito esforço.

5.2. TRABALHOS FUTUROS

O presente trabalho apresenta perspectiva futura em diversas áreas. O aprimoramento das técnicas e métodos desenvolvidos e a introdução de novas funcionalidades tendem a suprir as limitações destacadas. A seguir, estão relacionadas algumas funcionalidades futuras para o motor.

- **Comparativo com outras Abordagens de Classificação:** Comparar a abordagem utilizada com outras técnicas de recuperação como Índice Invertido por exemplo como forma de avaliar os resultados;

- **Adição de Mais Bases de Dados:** Vincular o motor de busca a mais bases de dados com intuito de enriquecer a base de dados com conteúdos;

- **Lematização Automática:** A partir dos parâmetros recebidos pelas aplicações, realizar integração com ferramentas de lematização de forma a utilizar todos lemas dos termos no momento da busca por conteúdos;

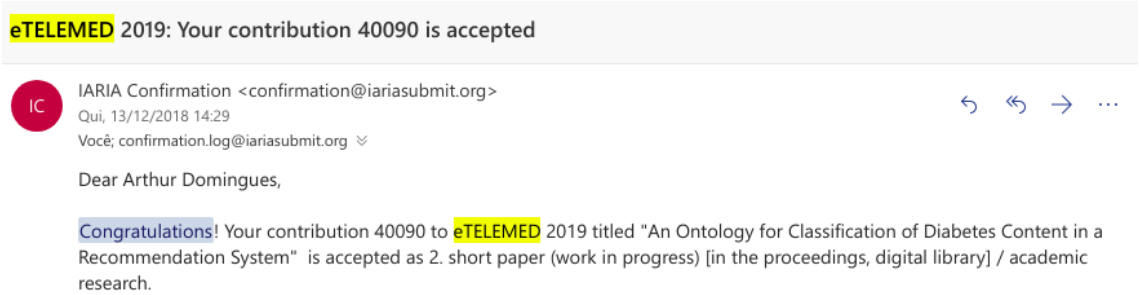
- **Técnicas de PLN:** Utilizar técnicas e algoritmos voltados para processo de linguagem natural (PLN) para leitura completa dos conteúdos textuais e não somente dos metadados;

- **Tradução de Documentos:** Utilizar tradutores automáticos a fim de disponibilizar conteúdos em idiomas como inglês e espanhol.

5.3. PRODUÇÃO CIENTÍFICA

Ao decorrer do desenvolvimento deste trabalho, as suas etapas bem como os resultados obtidos foram submetidos a conferências, eventos e periódicos:

- **Título**: An Ontology for Classification of Diabetes Content in a Recommendation System;
- **Veiculo**: eTELEMED: International Conference on eHealth, Telemedicine, and Social Medicine;
- **Situação**: Aceito para Publicação;



- **Título**: Semantic Data Integration Service for eHealth Applications;
- **Veiculo**: The 17th World Congress on Medical and Health Informatics;
- **Situação**: Aceito para Publicação;

MEDINFO 2019
HEALTH AND WELLBEING E-NETWORKS FOR ALL www.medinfo-lyon.org
CONVENTION CENTRE 26-30 AUGUST 2019, LYON FRANCE

Dear Mr. Graduate Domingues,

Decision: accepted as Poster

Congratulations: Your submission Semantic Data Integration Service for eHealth Applications (Id=781) was accepted as Poster by the SPC.
Please revise your manuscript according to the comments from the reviewers and the editorial recommendations provided below and **re-submit your revised manuscript by April 1st, 2019.**

- **Título**: Semantic Search Engines for Health Data: a Systematic Review;
- **Veiculo**: Revista IEEE Latin America;
- **Situação**: Revisões Necessárias;

[Transactions] New notification from IEEE Latin America Transactions

IC Ilse Cervantes <IlseCervantes@ieee.org>
Ter, 05/03/2019 16:23
Você

You have a new notification from IEEE Latin America Transactions:

You have been added to a discussion titled "Resubmit with modifications" regarding the submission "Semantic Search Engines for Health Data: a Systematic Review".

Link: https://www.inaoep.mx/~IEEElat/index.php/transactions/authorDashboard/submission/715

Alejandro Díaz

IEEE Latin America Transactions

6. REFERÊNCIAS

Agner, Luiz, and Anamaria Moraes. "Navegação e arquitetura de informação na web: a perspectiva do usuário." Boletim Técnico do Senac 29.1 (2018): 52-60.

Almeida, Maurício Barcellos. "Uma introdução ao XML, sua utilização na Internet e alguns conceitos complementares." Ciência da informação 31.2 (2002): 5-13.

Arya, Chandrakala, and Sanjay K. Dwivedi. "News web page classification using url content and structure attributes." Next Generation Computing Technologies (NGCT), 2016 2nd International Conference on. IEEE, 2016.

Baeza-Yates, Ricardo, and Berthier de Araújo Neto Ribeiro. Modern information retrieval. New York: ACM Press; Harlow, England: Addison-Wesley, 2011.

Barth, Fabrício Jailson. "Uma introdução ao tema Recuperação de Informações Textuais." Revista de Informática Teórica e Aplicada 20.2 (2013): 247-272.

Bax, Marcello Peixoto. "As bibliotecas na Web e vice-versa." Perspectivas em Ciência da Informação 3.1 (1998).

Bax, Marcello Peixoto. "Introdução às linguagens de marcas." Ciência da Informação 30.1 (2001): 32-38.

Biruel, E. P. (2008). Websites para diabéticos: uso da Internet como instrumento de Educação em Saúde. UNIFESP, São Paulo.

Brown, Heather. "Standards for structured documents." The Computer Journal 32.6 (1989): 505-514.

Bupa Helath Pulse. Disponível em: http://www.bupa.com/mediacentre/healthpulse. Acesso em 20 Agosto. 2018.

BUCKLAND, Michael; GEY, Fredric. The relationship between recall and precision. Journal of the American society for information science, v. 45, n. 1, p. 12-19, 1994.

Büttcher, S., Clarke, C. L., & Cormack, G. V. (2016). Information retrieval: Implementing and evaluating search engines. Mit Press, 2016.

Can, A. B., & Baykal, N. (2007). MedicoPort: A medical search engine for all. Computer methods and programs in biomedicine, 86(1), 73-86.

Cardoso, Olinda Nogueira Paes. "Recuperação de Informação." INFOCOMP 2.1 (2004): 33-38.

Chen, R. C., Huang, Y. H., Bau, C. T., & Chen, S. M. (2012). A recommendation system based on domain ontology and SWRL for anti-diabetic drugs selection. Expert Systems with Applications, 39(4), 3995-4006.

Cohen, Marcelo de Azevedo. Gerador de aplicações para consultas a bases RDF/RDFS. Diss. PUC-Rio, 2010.

Comitê Gestor da Internet no Brasil. TIC Domicílios e empresas 2012: pesquisa sobre o uso das tecnologias de informação e comunicação no Brasil. Junho 2013.

Costa, B. I. R. (2008). Coeficientes de revocação (recall) e precisão (precision) do sistema de recuperação de informação da biblioteca do ICEX da UFMG: usando amostra do acervo de teses e dissertações.

Croft, W. Bruce, Donald Metzler, and Trevor Strohman. Search engines: Information retrieval in practice. Vol. 520. Reading: Addison-Wesley, 2010.

Daconta, Michael C., Leo J. Obrst, and Kevin T. Smith. The Semantic Web: a guide to the future of XML, Web services, and knowledge management. John Wiley & Sons, 2003.

da Cunha Yamane, Gabriela Aparecida, and Fabiano Ferreira de Castro. "O estudo e a identificação dos padrões de metadados para a representação e a recuperação da imagem digital na perspectiva da web." Em Questão 24.1 (2018): 145-173.

Davanzo, Luciana. "Vocabulário controlado para arquivos: análise de viabilidade e propostas." Dissertação de Mestrado – Unesp (2016).

de Souza, Terezinha Batista, Maria Elizabete Catarino, and Paulo Cesar dos Santos. "Metadados: catalogando dados na Internet." Transinformação 9.2 (2012).

Desai, Keyur, et al. "Web Crawler: Review of Different Types of Web Crawler, Its Issues, Applications and Research Opportunities." International Journal of Advanced Research in Computer Science 8.3 (2017).

Dixit, Ashutosh, and A. K. Sharma. "A mathematical model for crawler revisit frequency." 2010 IEEE 2nd International Advance Computing Conference (IACC). IEEE, 2010.

Dragusin, R., Petcu, P., Lioma, C., Larsen, B., Jørgensen, H. L., Cox, I. J., ... & Winther, O. (2013). FindZebra: a search engine for rare diseases. International journal of medical informatics, 82(6), 528-538.

Ferguson, T., & Frydman, G. (2004). The first generation of e-patients: These new medical colleagues could provide sustainable healthcare solutions. BMJ: British Medical Journal, 328(7449), 1148.

Ferneda, Edberto. Introdução aos modelos computacionais de recuperação de informação. Ciência moderna, 2012.

Ferneda, Edberto; Zaira Regina Zafalon; Paula Regina Dalevedove(2017). “Ontologias na Representação e na Recuperação de Informação”. In: Perspectivas da Representação Documental: discussão e experiências. 1ed.São Carlos: CPOI/UFSCar, 2017, v. , p. 343-357.

Formenton, D., Ferreira de Castro, F., de Souza Gracioso, L., Furnival, A. C. M., & de Melo Simões, M. D. G. (2017). Os padrões de metadados como recursos tecnológicos para a garantia da preservação digital. Biblios, (68), 82-95.

Fox, Susannah. *Online health search 2006*. Pew Internet & American Life Project, 2006.

FRITZEN, Eduardo; SIQUEIRA, Sean WM; DE ANDRADE, Leila CV. Recuperação Contextual de Informação na Web para Apoiar Aprendizagem Colaborativa em Redes Sociais. In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). 2012.

Furgeri, Sérgio. "O papel das linguagens de marcação para a Ciência da Informação." TransInformação 18.3 (2006).

GARCIA, S. S. Metadados para documentação e recuperação de imagens. Dissertação (Mestrado) – Instituto Militar de engenharia (IME), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 1999, 138 p.

Garg, Abhinav, Kratika Gupta, and Abhijeet Singh. "Survey of Web Crawler Algorithms." International Journal of Advanced Research in Computer Science 8.5 (2017).

Gianotti, Priscila Salinas P., H. P. Pellegrino, and Elizabeth Wada. "Globalização e serviços médicos: impulsionando o turismo de saúde." Turydes [Internet] 4.2 (2009).

Goldfarb, C. F. (1990). The SGML handbook. Oxford University Press.

Gomes, Roberto Miranda. Desambiguação de Sentido de Palavras Dirigida por Técnicas de Agrupamento sob o Enfoque da Mineração de Textos. Diss. PUC-Rio, 2009.

Grácio, José Carlos Abbud. "Metadados para a descrição de recursos da Internet: o padrão Dublin Core, aplicações e a questão da interoperabilidade." (2002): 127-f.

Gruber, T. R. (1993). “A translation approach to portable ontology specifications”. Knowledge acquisition, 5(2), 199-220.

Guimarães, Célio. "Introdução a Linguagens de Marcação: HTML, XHTML, SGML, XML." (2003).

Gupta, S., & Bhatia, K. K. (2013, August). HiCrawl: A Hidden Web crawler for Medical Domain. In Computational and Business Intelligence (ISCBI), 2013 International Symposium on (pp. 152-157). IEEE.

Gómez-Pérez, Asunción. "Ontology evaluation." Handbook on ontologies. Springer, Berlin, Heidelberg, 2004. 251-273.

Harold, Elliotte Rusty, and X. M. L. Bible. "IDG Books Worldwide." (1999): 191-192.

Ide, N. C., Loane, R. F., & Demner-Fushman, D. (2007). Essie: a concept-based search engine for structured biomedical text. Journal of the American Medical Informatics Association, 14(3), 253-263.

Ikematu, Ricardo Shoiti. "Gestão de metadados: sua evolução na tecnologia da informação." DataGramaZero-Revista de Ciência da Informação 2.6 (2002).

Isotani, S., & Bittencourt, I. I. (2015). “Dados Abertos Conectados: Em busca da Web do Conhecimento”. Novatec Editora.

Kiryakov, A. (2006). “Ontologies for knowledge management”. Semantic Web Technologies: Trends and Research in Ontology-based Systems, 115-138

Lampesberger, H. (2016). Technologies for Web and cloud service interaction: a survey. Service Oriented Computing and Applications, 10(2), 71-110.

Lancaster, F. W. "Indexação e resumos: teoria e prática (Indexing and abstracts: theory and practice)." (2004).

Laufer, C. "Guia de Web Semântica." São Paulo: CeWeb-Centro de Estudos sobre (2015).

Li, M., Li, C., Wu, C., & Luo, Y. (2015). A focused crawler url analysis algorithm based on semantic content and link clustering in cloud environment. International Journal of Grid & Distributed Computing, 8(2).

Lin, Yu, and Norihiro Sakamoto. "Ontology driven modeling for the knowledge of genetic susceptibility to disease." Kobe J Med Sci 55.3 (2009): E53-E66.

Luo, G., Tang, C., Yang, H., & Wei, X. (2008, October). MedSearch: a specialized search engine for medical information retrieval. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 143-152). ACM.

Luo, G. (2009, March). Design and evaluation of the iMed intelligent medical search engine. In Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on (pp. 1379-1390). IEEE.

Marcondes, Carlos Henrique. "“Linked data”–dados interligados-e interoperabilidade entre arquivos, bibliotecas e museus na web." Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação 17.34 (2012): 171-192.

Martins, Aline Freitas. “Construção de ontologias de tarefa e sua reutilização na engenharia de requisitos”. 2009. 161 f. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal do Espírito Santo, Vitória, 2009.

McDaid, D., & Park, A. L. (2010). Online health: untangling the web.

Moreno, Fernanda Passini, and Marisa Brascher. "MARC, MARCXML e FRBR: relações encontradas na literatura." Informação & Sociedade: Estudos 17.3 (2007).

Murray, E., Burns, J., See Tai, S., Lai, R., & Nazareth, I. (2005). Interactive Health Communication Applications for people with chronic disease. The Cochrane Library.

Nardon, Fabinane Bizinella. "Utilizando XML para a representação de informação em saúde." Newsletter SBIS (2000).

Noy, Natalya F., and Deborah L. McGuinness. "Ontology development 101: A guide to creating your first ontology." (2001).

Pereira, Ana Maria, Divino Ignácio Ribeiro Júnior, and Guilherme Luiz Cintra Neves. "Metadados para a descrição de recursos da internet: as novas tecnologias desenvolvidas para o padrão Dublin Core e sua utilização Metadata for internet resources description: new technologies developed for the Dublin Core standard and its use p. 7-39." Revista ACB 10.1 (2005): 7-39.

Pessoni, Arquimedes. "Mergulho no mar de informações da web." (2012): 453-455.

Ramirez, D., Lago, P., Borda, D., & Jiménez-Guarín, C. (2011, May). Mental health information Web search and semantic search extension. In Computing Congress (CCC), 2011 6th Colombian (pp. 1-6). IEEE.

Ribeiro, Gilberto Pessanha. "Metadados geoespaciais digitais." WORKSHOP DE BANCOS DE DADOS NÃO CONVENCIONAIS,(2.: 1995: Niterói) Anais... Niterói. 1995.

Ribeiro-Neto, Berthier, and Ricardo Baeza-Yates. "Modern information retrieval." New York: ACM Press, Addison-Wesley (1999).

Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. Journal of documentation, 33(2), 106-119.

Rosa, Paulo Augusto. Web Semântica. Diss. Tese (Doutorado)—Universidade de São Paulo, Instituto de Matemática e Estatística, 2002.

Rossi, Ricardo Messias. Caracterização e coordenação de sistemas produtivos: o caso do trigo no Brasil. Diss. Universidade de São Paulo, 2004.

Santos Neto, Martins Fideles dos. (2013). ONTOLIME: Modelo de Ontologia Descrição e Imagens Médicas 207 f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista “Júlio de Mesquita Filho”.
Sarrouti, M., & El Alaoui, S. O. (2017). A Yes/No Answer Generator Based on Sentiment-Word Scores in Biomedical Question Answering. International Journal of Healthcare Information Systems and Informatics (IJHISI), 12(3), 62-74.

Sayão, Luís Fernando. "Padrões para bibliotecas digitais abertas e interoperáveis 10.5007/1518-2924.2007 v12nesp1p18." Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação 12.1 (2007): 18-47.

Selvakumar, M., Prabhu, D., & Kathiravan, A. V. (2012). A novel Community based Web Crawlers (CWC) for Information Retreival. International Journal of Computer Applications, 48(20), 29-31.

Silva, Fernando Ferraz. "Repositórios institucionais: sugestão de metadados para diversos suportes informacionais." (2008).

Silva, Renata Eleuterio da, Plácida Leopoldina Ventura Amorim da Santos, and Edberto Ferneda. "Modelos de recuperação de informação e web semântica: a questão da relevância." Informação & Informação (2013): 27-44.

Soares, M. C. (2004). Internet e saúde: possibilidades e limitações. Textos de la Ciber Sociedad, 4.

Souza, A. S., Duran, A., & Vieira, V. (2014). “Um estudo de mapeamento sistemático sobre ontologias para a metodologia de aprendizagem baseada em problemas”. In Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE) (Vol. 25, No. 1, p. 1103).

Souza, Marcia Isabel Fugisawa, Laurimar Gonçalves Vendrusculo, and Geane Cristina Melo. "Metadados para a descrição de recursos de informação eletrônica: utilização do padrão Dublin Core." Ciência da Informação 29.1 (2000).

Sombra, E. L. (2015). MobiLEHealth: Um ambiente de Apoio à Saúde 2.0. Universidade do Estado do Rio Grande do Norte-UERN e Universidade Federal Rural do Semi-Árido-UFERSA, Mossoró-RN.

Trivedi, M. (2009). A study of search engines for health sciences. International Journal of Library and Information Science, 62(68), 062-073.

Uschold, M., & Gruninger, M. (1996). “Ontologies: Principles, methods and applications”. The knowledge engineering review, 11(2), 93-136.

Vega, Arturo Martín. Fuentes de información general. Trea, 1995.

Weitzel, L., de Oliveira, J. P. M., Boito, F. Z., dos Santos, H. D. P., Nobre, J. C., Lutz, J. A. F., ... & Moraes, T. G. Proposta de métricas de avaliação da qualidade da informação médica para Sistemas de Recomendação baseados no perfil do usuário. Cadernos de Informática, 5(1), 23-48, 2010.

ANEXOS

Questionário de Avaliação de Conteúdo

Este questionário é apenas para fins acadêmicos, o objetivo é analisar o nível de satisfação dos usuários sobre os conteúdos científicos recomendados pelo sistema MobiLEHealth. Por gentileza responda a todas as perguntas. Obrigado por sua ajuda e atenção.

*Obrigatório

1. **Faixa Etária** *
   Marcar apenas uma oval.

   - 18 - 24 Anos
   - 25 - 30 Anos
   - 31 - 35 Anos
   - 36 - 40 Anos
   - Mais de 40 Anos

2. **Quantos conteúdos você leu?** * *Marcar apenas uma oval.*

   - Entre 1 e 3 Conteúdos
   - Entre 3 e 5 Conteúdos
   - Entre 5 e 7 Conteúdos

3. **Os conteúdos recomendados apresentam termos referentes ao diabetes?** * *Marcar apenas uma oval.*

   - 1 - Discordo Totalmente
   - 2 - Discordo Parcialmente
   - 3 - Não Concordo, Nem Discordo
   - 4 - Concordo Parcialmente
   - 5 - Concordo Totalmente

4. **Quanto aos temas dos conteúdos, você os considera, de um modo geral, interessantes?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

5. **Você concorda com o nível de recomendação atribuído a cada conteúdo?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

6. **Na sua opinião, os conteúdos recomendados podem auxiliar no tratamento do diabetes?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

7. **Na sua opinião, os conteúdos recomendados permitem que os usuários obtenham mais clareza quanto ao Diabetes?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

8. **Você considera que os conteúdos recomendados são adequados em relação ao Diabetes?**
*

Marcar apenas uma oval.

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

9. **Você considera que o número de conteúdos recomendados é satisfatório?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

10. **Você indicaria os conteúdos recomendados para um paciente e/ou profissional de saúde?** *

Marcar apenas uma oval.

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente
- 5 - Concordo Totalmente

11. **Em geral, você considera que os conteúdos recomendados proporcionam uma melhor qualidade de vida aos pacientes diabéticos?** * *Marcar apenas uma oval.*

- 1 - Discordo Totalmente
- 2 - Discordo Parcialmente
- 3 - Não Concordo, Nem Discordo
- 4 - Concordo Parcialmente