Josceli Maria Tenório

VOCABULÁRIO DE SAÚDE DO CONSUMIDOR EM IDIOMA PORTUGUÊS

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina, para defesa de tese no curso de Doutorado do Programa de Pós-graduação em Gestão e Informática em Saúde.

São Paulo
2019

Josceli Maria Tenorio

VOCABULÁRIO DE SAÚDE DO CONSUMIDOR EM IDIOMA PORTUGUÊS

Tese apresentada à Universidade Federal de São Paulo – Escola Paulista de Medicina, para defesa de tese no curso de Doutorado do Programa de Pós-graduação em Gestão e Informática em Saúde.

Orientador:
Prof. Dr. Ivan Torres Pisa

São Paulo
2019

Tenório, Josceli Maria
Vocabulário de saúde do consumidor em idioma português / Josceli Maria Tenório. -- São Paulo, 2019.
xxv, 131 f.

Tese – Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós-Graduação em Gestão e Informática em Saúde.

Título em inglês: Consumer health vocabulary in Brazilian Portuguese language

1.Vocabulário Controlado. 2. Vocabulário. 3. Informática Aplicada à Saúde dos Consumidores. 4.Informação de Saúde ao Consumidor.

Universidade Federal de São Paulo

Escola Paulista de Medicina

Programa de Pós-Graduação em Gestão e Informática em Saúde

Coordenadora da Câmara de Pós-graduação e Pesquisa da Escola Paulista de Medicina:

Profa. Dra. Mônica Levy Andersen, livre docente

Coordenador do programa:

Prof. Dr. Ivan Torres Pisa, livre docente

Josceli Maria Tenorio

VOCABULÁRIO DE SAÚDE DO CONSUMIDOR EM IDIOMA PORTUGUÊS

Presidente da banca:

Prof. Dr. Ivan Torres Pisa, livre docente (orientador)
Universidade Federal de São Paulo, Escola Paulista de Medicina, Departamento de Informática em Saúde.

Banca examinadora:

Prof. Dr. Evandro Eduardo Seron Ruiz
Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto, Departamento de Computação e Matemática.

Prof. Dr. Paulo Bandiera Paiva
Universidade Federal de São Paulo, Escola Paulista de Medicina, Departamento de Informática em Saúde.

Dr. Roberto Silva
Associação Saúde da Família, São Paulo.

Prof. Dr. Stefan Schulz
Institute of Medical Informatics, Statistics and Documentation, General Hospital and University Clinics, Austria.

Suplentes:

Profa. Dra. Claudia Galindo Novoa

Universidade Federal de São Paulo, Escola Paulista de Medicina, Departamento de Informática em Saúde.

Prof. Dr. Paulo Mazzoncini de Azevedo Marques

Universidade de São Paulo, Faculdade de Medicina de Ribeirão Preto, Departamento de Clínica Médica.

Dedicatória

À minha mãe Ewa,
o sentido da minha vida.
Ao meu amado pai,
Rômulo Batista de Assis (*in memorian*).
Às minhas amadas mães
Maria José (*in memorian*), Anas,
Marias, Senhora, Márcia.

Agradecimentos

Este trabalho é um dos frutos do esforço, aprendizado e conhecimento de muitas pessoas. Minha homenagem a todos os que contribuíram com esta realização. Nominar estas pessoas é muito difícil, porque, às vezes, apenas uma pergunta simples pode ter o potencial de mudar uma visão ou de resolver problemas.

Meus sinceros agradecimentos ao Prof. Dr. Ivan Torres Pisa, orientador desta tese, pelo aprendizado diário, amizade, oportunidades, paciência, compreensão, boas conversas e, principalmente, por me apoiar nos momentos mais difíceis da minha vida, que ocorreram durante o desenvolvimento desse projeto.

Aos membros das bancas de projeto e qualificação, por todas as críticas, esclarecimentos, contribuições e disponibilidade, essenciais para a melhoria e finalização desse trabalho, meus agradecimentos.

Agradeço aos membros da banca de defesa de tese, pelo aceite do convite e pela disponibilidade em colaborar com este trabalho, uma honra para mim.

A todos os professores do Programa de Pós Graduação em Gestão e Informática em Saúde (PPGIS) – Escola Paulista de Medicina – UNIFESP e ao Departamento de Informática em Saúde (DIS). Meus agradecimentos ao Prof. Dr. Meide Silva Anção, chefe do departamento e à secretária do PPGIS Valdice P. dos S. Ribeiro.

Obrigada a todos os participantes do grupo de pesquisa Saúde 360°, aos quais devo aprendizagem e a oportunidade de poder discutir ideias, meus colegas e amigos que compartilharam diversos momentos durante esses anos. Mesmo sob o risco de cometer injustiça, não posso me furtar a agradecer nominalmente a algumas pessoas: Fernando Sequeira (que iniciou este projeto), Gabriela, Roberto Baptista, Siony e Carolina Viviani (por todo apoio na avaliação de termos), Thiago, Gilberto, Frederico Cohrs, Leonardo, minha irmã Simone Vitti e Paulo Lopes. Agradeço especialmente ao Adalberto Tardelli e ao Fabrício Landi, que disponibilizou a infraestrutura para o desenvolvimento desse trabalho.

Aos meus colegas e amigos do Departamento de Informática e Turismo - Instituto Federal de São Paulo, que torceram por mim e me apoiaram em momentos difíceis durante este projeto. Meu agradecimento especial ao diretor César Lopes

Fernandes, a todos os coordenadores e a minha querida Terezinha.

À minha família, que compartilhou e torceu por mim em toda minha vida.

A todos os meus amados amigos e irmãos, em especial, Shiro, Mara e Marcos, que sempre estiveram comigo, partilhando alegrias e dificuldades.

À minha amada mãe, Senhora, e a todos os meus amados irmãos de Salvador, que me acolheram num momento tão difícil da minha vida. Obrigada por tudo.

À minha mãe, Márcia, que me presenteou com palavras maravilhosas e me tirou da escuridão, todo meu respeito, amor, gratidão e admiração.

A todos os colegas e profissionais atuantes no DIS, por todos os bons momentos de amizade e companheirismo, materializados pelo nosso café e por boas risadas.

Sumário

Lista de figuras

Lista de tabelas

Lista de quadros

Lista de abreviaturas e símbolos

CHV	Consumer Health Vocabulary
CID-10	Classificação Internacional de Doenças
CUI	UMLS Concept Unique Identifier
DeCS	Descritores em Ciência da Saúde
DUI	MeSH Descriptor Unique Identifier
HON	Health on the Net Foundation
HTML	HyperText Markup Language
IRI	Internationalized Resource Identifiers
MeSH	Medical Subject Headings
NER	Named Entities Recognition
NLM	National Library of Medicine
OAC-CHV	Open Access Collaborative Consumer Health Vocabulary Initiative
OWL	Web Ontology Language
PAV	Provenance, Authoring and Versioning
POS	Part of Speech
PROV	PROVenance Interchange Ontology
RAKE	Rapid Algorithim Keyword Extraction
RAT	Reconhecimento Automático de Termos
RDF	Resource Description Framework
RSS	Really Simple Syndication
SKOS	Simple Knowledge Organization System
SNOMED	Systematized Nomenclature of Medicine
SQL	Structured Query Language
SPARQL	Simple Protocol and RDF Query Language
UMLS	Unified Medical Language System
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VSC	Vocabulário de Saúde do Consumdor
W3C	World Wide Web Consortium
XML	Extensible Markup Language

Lista de publicações

Artigo publicado

Tenório JM, Pisa IT. Consumer Health Vocabulary: A proposal for a Brazilian Portuguese language. Stud Health Technol Inform. 2015;216:1089.

Artigo aceito

Tenório JM, Pisa IT. Identificação automática de termos de domínio do consumidor em saúde. Journal of Health Informatics. Março/2019. No prelo.

Artigo submetido

Tenório JM, Landi-Moraes F, Pisa IT. CHV.br: Exploratory study for development of a consumer health vocabulary (CHV) supported by a network model for Brazilian Portuguese language. Health Informatics Journal. Setembro/2019.

Artigos em elaboração

Tenório JM, Landi-Moraes F, Viviani CM, Pisa IT. Identificação de termos de saúde do consumidor baseada em uma abordagem não supervisionada.

Tenório JM, Landi-Moraes F, Pisa IT. CHV.br-RDF: potencialidades para o enriquecimento de dados do consumidor em saúde.

Apoio financeiro

Este projeto recebeu apoio financeiro por meio da concessão de bolsa de doutorado da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001, período março/2015 a fevereiro/2019.

Tenorio JM. Vocabulário de saúde do consumidor em idioma português [tese – Doutorado]. São Paulo. Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós-Graduação em Gestão e Informática em Saúde; 2019. 131f.

Resumo

Introdução: Estudos mostram lacunas entre os termos usados por consumidores e termos técnicos usados por profissionais de saúde. Os vocabulários de saúde do consumidor (VSC) são apresentados como uma solução em especial incorporam tecnologias que possibilitam a disponibilização e integração de conteúdo e de relações semânticas entre os termos. **Objetivo:** Desenvolver um VSC em idioma português baseado em conteúdos disponíveis na web e estruturado segundo os princípios e tecnologias da web semântica. **Método:** Este estudo foi dividido em três fases. Na Fase 1 foram coletados e extraídos termos de conteúdos estruturados disponíveis na web, como os vocabulários controlados do Unified Medical Language System (UMLS) e a base de conhecimento DBpedia. Esses termos e suas relações semânticas foram representados por meio de rede complexa. Medidas de centralidade foram efetuadas para caracterização da rede. A seleção dos termos para compor o VSC foi realizada por meio da aplicação de técnicas de clusterização da rede. A Fase 2 foi composta por duas estratégias aplicadas para incluir termos não obtidos na Fase 1 a partir da análise de conteúdos não estruturados escritos por ou para consumidores em saúde, sendo a recuperação de termos dos vocabulários do UMLS e a utilização de técnicas para reconhecimento automático para identificar termos candidatos. Um processo de validação humana foi aplicado para homologar os termos candidatos para inserção no VSC. Na Fase 3 o conteúdo do VSC foi formalizado por ontologias específicas e representado pelo modelo Resource Description Framework (RDF). Uma interface web foi construída para acesso aos dados. **Resultados:** A Fase 1 resultou em uma rede complexa composta por 146.956 termos ligados por relações semânticas como sinonímia, hiperonímia, além de termos relacionados, dos quais 31.439 são conceitos do UMLS representados por termos preferidos e 83.279 são sinônimos. A DBpedia contribuiu com a elevação

da taxa sinônimo/conceito de 1,6 para 2,5. Medidas de centralidade mostraram características da rede revelando termos significativos. A Fase 2 resultou na recuperação de 5.916 termos dos vocabulários do UMLS. O algoritmo para reconhecimento automático de termos resultou na obtenção de 9.674 n-gramas candidatos. A validação humana resultou em 6.245 termos admissíveis ao VSC (66,24% dos termos avaliados). A curva de precisão-revocação do algoritmo resultou em [0,732-~0,900], um valor superior ao apresentado em estudos análogos. Na Fase 3 os dados foram formalizados usando as ontologias Simple Knowledge Organization System (SKOS) e Provenance, Authoring and Versioning (PAV) que mostraram-se adequadas para formalizar o conteúdo do VSC e suportar o modelo RDF. O modelo VSC-RDF foi construído com 150.995 termos dos quais 66.992 são termos preferidos e 84.003 sinônimos, além do mapeamento de outros relacionamentos semânticos baseados em hierarquia e associação. **Conclusão:** A partir dos dados disponibilizados na web foi possível construir um modelo de VSC por meio da aplicação de técnicas computacionais. O modelo de rede possibilita ligar termos provenientes de vocabulários controlados e do consumidor, representar seus relacionamentos semânticos e foi a base para a construção do modelo VSC-RDF. Sinônimos, termos e relações inéditos foram identificados. Este estudo apresentou uma infraestrutura de dados para o desenvolvimento de aplicações orientadas ao consumidor e propõe um método de desenvolvimento de vocabulários de saúde em outros idiomas e de atualização de vocabulários existentes.

Abstract

Introduction: Some research studies show a distant language gap between the common terms used by laypersons and the technical terms used by healthcare professionals. Thus, a proposed solution to this language gap barrier is the consumer health vocabularies (CHV) index, where could be incorporated technologies which makes the data available, integrated as well as semantic relationships between themselves. **Objective:** Developing a Brazilian Portuguese CHV model based on web data sources, and structured according to semantic web vocabulary principles and technologies. **Method:** This study was split into three distinct phases. In Phase1, we have collected and extracted terms from some web-structured data sources, such as the Unified Medical Language System (UMLS) controlled vocabularies and the DBpedia Knowledge Base. These terms and their semantic relationships have represented by a complex network. Some network centrality measures have been obtained in order to characterise it. The selection of terms which could compose the CHV was performed through clustering network techniques. Phase 2 was conducted based on two steps in order to obtain new terms from unstructured web data sources written by and/or for health consumers, composed by recognition of UMLS' terms and use of term automatic recognition techniques in order to identify candidate terms. A human validation process was conducted in order to approve these candidate terms and insert them into the CHV. In Phase 3 the CHV data have formalised through specific ontologies and have represented by the Resource Description Framework (RDF) web data model. Furthermore, we designed and developed a layout to access the dataset by users. **Results:** Phase 1 resulted into a complex network containing already 146,956 terms linked by semantic relationships as synonyms, hyperonymy, and related terms, of which 31,439 are UMLS concepts, represented by preferred terms and 83,279 are synonyms. DBpedia have raised the synonym per concept rate from 1.6 to 2.5. Centrality measures were important to show some characteristics of the complex network in order to reveal the most important terms. Phase 2 has resulted in the automatic recognition of 5,916 UMLS terms. The term automatic recognition algorithm allowed recognizes 9,674 n-grams candidates. Human validation has validated 6,245 terms, around 66.24% of the

candidate terms assessed. The precision-recall curve of the algorithm that performed the automatic term recognition resulted in [0.732- ~ 0.900], a greater value than founded by other similar studies. In Phase 3, we formalized these data using Simple Knowledge Organization System (SKOS) and Provenance, Authoring and Versioning (PAV), suitable ontologies for CHV and supporting RDF data model. The CHV-RDF contains already 150,995 terms, which of 66,992 are preferred terms, and 84,003 are synonyms, besides the mapping of other semantic relationships between terms based on hierarchy and association. **Conclusion:** It was possible to build a CHV model automatically through computational techniques using data sources available on the web. The complex network model enabled to link and match terms provided by controlled and consumer vocabularies, represent their semantic relationships, and it has supported the CHV-RDF data model. Unpublished synonyms, terms and relationships have been identified. This study showed a data infrastructure which could be used for the development of consumer-oriented applications and proposed a method to development of health vocabularies in other language and updating existing vocabularies.

1 INTRODUÇÃO

A comunicação entre consumidores de informação em saúde e profissionais de saúde é um item potencial para trazer benefícios aos tratamentos clínicos e à permanência do indivíduo em estado saudável. Entretanto, falhas de comunicação são comuns entre as partes e podem ocorrer devido ao distanciamento entre os vocabulários utilizados por pacientes ou consumidores em saúde e por profissionais de saúde.[1, 2] Uma das consequências dessa falha é a dificuldade na localização de informações em saúde em uma busca na web, que passa a ser uma questão relevante devido ao aumento expressivo de pesquisas realizadas.[3]

Consumidores que buscam informações sobre saúde e doenças na web para fins diversos têm sido foco de estudo desde o início dos anos 2000.[3, 4] Questões relacionadas à saúde compõem a segunda área temática mais pesquisada do Google, compondo 5% das mais de dois trilhões de pesquisas realizadas em 2016.[5] Estudos recentes mapearam o comportamento dos consumidores ou pacientes que buscaram informações na web e mostraram que esses são jovens com boa literacia médica e que os efeitos positivos constam do impacto positivo na interação médico-paciente [5] e desenvolvimento de habilidades como entender, avaliar ou sintetizar, necessárias a utilização da informação em saúde.[6]

Apesar dos benefícios apontados por estes estudos, há uma preocupação referente à atualização e confiabilidade das informações recuperadas [7], as quais nem sempre são seguidas pelas fontes de informação consultadas.[8] A estratégia dos consumidores para encontrar respostas para questões de saúde consta de utilizar vários termos de pesquisa, explorando os primeiros resultados por exame superficial do conteúdo da página e refinando iterativamente sua estratégia de pesquisa.[4] Isso sugere que os consumidores têm seu foco na utilização de termos e este pode ser um caminho potencial para o desenvolvimento de aplicações para o consumidor.

Uma possibilidade para atender a demanda por informação é desenvolver ferramentas que auxiliem esses consumidores em sua interação com os conteúdos de saúde retornados pelos buscadores e com outras ferramentas que proveem conteúdos de saúde de forma a aperfeiçoar o processo de busca por informações

realizada pelos consumidores. Uma iniciativa desenvolvida nos últimos anos é a criação de vocabulários controlados com foco nos consumidores, conhecidos sob a denominação de Consumer Health Vocabulary (CHV).[9, 10] A essência do CHV é a possibilidade de suportarem o desenvolvimento de aplicações para o consumidor por meio da ligação de termos leigos a termos técnicos.

Questões referentes à formalização das definições que suportam este estudo, a exemplo de termo, conceito, termo técnico e leigo, serão tratadas no Capítulo 2 desta tese.

Nas próximas seções serão discutidos os conceitos referentes ao desenvolvimento de CHVs.

1.1 Desenvolvimento de CHVs

Os CHVs são vocabulários controlados que segundo Zeng [9, 10] "ligam termos cotidianos sobre saúde a termos técnicos ou jargões utilizados por profissionais de saúde".[11]

A importância de se criar os CHVs é o fato de haver uma distância entre o vocabulário médico, com termos mais técnicos e de pouca difusão entre consumidores de saúde, e o vocabulário propriamente utilizado pelos próprios consumidores. Os CHVs devem se manter constantemente atualizados devido à dinamicidade dos idiomas e surgimento de novos termos e conceitos de saúde.

Para o idioma inglês existe um CHV consolidado, suportado pela National Library of Medicine (NLM) (nlm.nih.gov) e desenvolvido segundo a estrutura proposta para vocabulários da UMLS.[12] Este vocabulário conta com um mapeamento dos termos voltados para o público em geral com os termos presentes no Unified Medical Language System (UMLS) [9, 13, 14], composto majoritariamente por termos técnicos da área médica e de saúde. Esse CHV conta com uma iniciativa para acesso e manutenção dos termos por meio de colaboração via ambiente wiki, a Open Access Collaborative Consumer Health Vocabulary Initiative (OAC-CHV) [11], que ficou disponível para acesso até julho de 2018. A OAC-CHV tem sido a base para apoiar estudos direcionados a sua manutenção e atualização automática [15] e para tentar superar desafios e problemas, como a recuperação automática de

termos para fins de atualização do OAC-CHV.[16] No presente estudo este vocabulário é identificado como OAC-CHV.

Com relação à construção de vocabulários, a técnica mais simples consta de registrar diretamente com consumidores de saúde os termos utilizados com frequência para se referir a situações de saúde. Outra possibilidade seria a tradução dos termos da OAC-CHV para o idioma português a exemplo dos Descritores em Ciências da Saúde DeCS (decs.bvs.br). Porém, devido à grande quantidade de conceitos existentes a verificação manual torna-se trabalhosa e inviável.[15] Outro fator a ser considerado é que este método não propiciaria encontrar novos termos do domínio da saúde exclusivos do idioma português e próprios do sistema nacional de saúde.

Técnicas de mineração de textos têm sido utilizadas com sucesso para a recuperação automática de termos para compor vocabulários.[17] A aplicação dessas técnicas possibilita aproveitar a vasta quantidade de informação disponível na web, como textos de notícias voltadas para a população e mensagens de redes sociais e fóruns, postadas majoritariamente por pessoas leigas em saúde, para encontrar termos candidatos para um vocabulário.[17] As contribuições do consumidor para a construção de vocabulários controlados parecem ser seriamente sub pesquisadas dentro e fora dos cuidados de saúde.[18] Pesquisadores de saúde precisam se envolver com os consumidores para garantir que as informações podem realmente suportar a comunicação com os profissionais de saúde.

A utilização de conteúdo eletrônico disponível na web também possibilita que novas tecnologias sejam aplicadas quanto à forma de disponibilização do conteúdo. As opções mais promissoras são as tecnologias que suportam a web semântica.[19]

A web semântica preconiza a utilização de tecnologias para representação, integração, disponibilização, conexão e reutilização dos dados. Fornece uma estrutura comum que possibilita que os dados sejam compartilhados e reutilizados segundo um modelo de intercâmbio de dados na web, o Resource Description Framework (RDF).[20] Nessa perspectiva os vocabulários são constituídos por conceitos e relacionamentos usados para descrever e representar uma área de estudo.[21]

As tecnologias da web semântica possibilitam a criação de vocabulários e das

regras para gerenciamento e disponibilização, assim como a utilização de tecnologias para conexão de dados, de forma que esteja em formato legível para máquinas.[20] O conceito de vocabulário proposto pela web semântica suporta a expansão do relacionamento de equivalência entre termos, ou sinonímia, usual entre vocabulários simples, no sentido de aumentar a abrangência das relações entre os conceitos, sejam hierárquicas ou associativas, como nos tesauros, por meio de relações exclusivas do domínio, a exemplo de “vírus (termo)” causa (relação) “influenza (termo)”.

O mapeamento de termos de domínio dos consumidores em saúde pode ser um item potencial para descobertas sobre os conceitos, suas relações e utilização. Para esta finalidade este estudo explora o caráter prático referente à construção de vocabulários. Os aspectos conceituais, pertencentes ao domínio da linguística, são tratados apenas de forma suficiente para a compreensão da perspectiva apresentada.

1.2 Relevância dos CHVs

A construção de bases de dados de domínios definidos é importante para fins de organização do conhecimento por suportarem o desenvolvimento de aplicações diversas. A possibilidade de desenvolvimento de aplicações é um resultado importante da iniciativa OAC-CHV.[9] Diversos itens utilizados por consumidores ou pacientes, como materiais educativos, rótulos de medicamentos, formulários, prontuários pessoais, grupos de discussão etc., requerem o conhecimento adequado de vocabulário.[2]

As aplicações eletrônicas requerem o processamento da informação de forma automatizada. Os sistemas de computador não podem reconhecer os termos utilizados pelos consumidores se não estiverem incluídos em bases de dados eletrônicas. Diante da disponibilidade de dados em diversas bases na web e das tecnologias desenvolvidas para troca de dados, uma possibilidade para a construção de vocabulários de domínio específico é a extração automática de conteúdo utilizando bases estruturadas como os dicionários multilinguagens e multidomínios do UMLS, OAC-CHV ou mesmo conteúdos textuais. Essas fontes possibilitam extrair

os termos e relações estabelecidas por eles, além de capturar termos exclusivos, de acordo com o idioma e sistema de saúde. A organização e disponibilização desses conteúdos são a base para o desenvolvimento de aplicações computacionais mais abrangentes.

Outra questão que envolve o desenvolvimento de vocabulários é o idioma. Nos últimos 15 anos as iniciativas de desenvolvimento de vocabulários ocorreram quase exclusivamente para o idioma inglês.[17] Em 2018 um estudo foi publicado descrevendo um vocabulário com termos relacionados exclusivamente a câncer em idioma francês.[22] Não foi encontrado nenhum trabalho publicado internacionalmente que descrevesse a construção de vocabulários de saúde do consumidor para o idioma português na literatura disponível no período citado.

A partir da sua versão inicial, disponibilizada na web, a OAC-CHV tem sido utilizada para apoiar estudos diversos. Métodos baseados em análise de textos recuperados da Wikipédia (wikipedia.org) ou de mídia social têm sido utilizados para fins de atualização do OAC-CHV, prospectando novos termos ou sinônimos.[16, 23] Tais estudos são restritos a obtenção de novos termos, porém não mostram que outras relações são estabelecidas entre termos, que poderiam ser a base para descoberta de fenômenos.

Em relação à utilização das tecnologias da web semântica, em 2016 foi publicada uma ontologia para o OAC-CHV.[24] Entretanto, não foram encontrados estudos que mostrassem sua evolução com o objetivo de promover o desenvolvimento de aplicações que utilizem o CHV no modelo para intercâmbio de dados RDF, o que possibilitaria a ligação entre bases de conhecimento para fins de enriquecimento de dados.

1.3 Outras iniciativas para o desenvolvimento de CHVs

Soluções para o problema de desenvolver aplicações projetadas para consumidores em saúde incluem a criação de vocabulários médicos, terminologias e ontologias. O desafio é mapear e utilizar recursos tecnológicos disponíveis.

Os estudos de Zeng [9, 10] focados na construção da primeira versão da OAC-CHV, iniciada em 2006, foram importantes para reforçar a necessidade de criar

estruturas para suportar aplicações orientadas a preencher a lacuna terminológica.

Um dos primeiros esforços para a organização de uma terminologia de saúde orientada para o consumidor foi o Multilingual Glossary of Technical and Popular Medical Terms in Nine European Languages [25], iniciado em 2000. Este vocabulário contém cerca de 1.400 termos leigos e técnicos e foi disponibilizado em nove idiomas europeus, inclusive português.

O Mayo Consumer Health Vocabulary, lançado em 2013, é outro esforço para a obtenção de uma taxonomia para o consumidor e possui aproximadamente 5.000 termos de saúde do domínio do consumidor, parcialmente mapeado para os vocabulários controlados Classificação Internacional de Doenças (CID-10) ou o Systematized Nomenclature of Medicine (SNOMED).[26] Os autores aplicaram técnicas de mineração de textos para expandir sua cobertura, o que possibilitou integrar termos da UMLS classificados como “doenças” a fatores genéticos e de risco para esses termos.

Em 2018 Tapi Nzali et al.[22] descreveram a construção de um vocabulário, em idioma francês, exclusivo para termos referentes a câncer de mama. Esse modelo foi construído por meio da aplicação de técnicas de extração de termos e identificação de relações semânticas, porém, restrita ao mapeamento de sinônimos. Este estudo propôs uma ontologia para a formalização do conteúdo. No mesmo ano Hou et al.[27] descreveram o desenvolvimento inicial de um CHV baseado no processo de extração de termos de saúde usados por consumidores chineses a partir de fóruns online e conteúdos textuais escritos para o consumidor em saúde.

A Wikipedia (wikipedia.org), lançada em 2001, tem sido uma importante fonte de dados provenientes dos usuários. A DBpedia (wiki.dbpedia.org/) é o projeto que visa extrair conteúdo estruturado da Wikipedia. Mrabet et al.[28] utilizaram em seus experimentos a combinação dos vocabulários do UMLS e DBpedia para recuperar termos em conteúdos referentes a questões escritas por consumidores em saúde e, numa segunda etapa, mapearam a classificação (tópico) de cada termo, de forma a obter um conjunto de termos pertencentes a tópicos específicos, como “doenças”. Os autores relatam uma melhora substancial na obtenção do tópico ao qual o termo pertence ao realizar o experimento combinando as bases. Tilahun et al.[29] utilizaram a DBpedia, vocabulários do UMLS e outras fontes como um meio para o

enriquecimento de dados referentes a HIV utilizando tecnologias da web semântica.

A extensão da cobertura da OAC-CHV foi base de vários estudos que aplicaram métodos baseados na extração automática de termos a partir de textos escritos por consumidores em saúde. Zeng et al.[30] compararam técnicas para a identificação automática de termos. Doing-Harris et al.[16] propuseram a utilização de uma rede social específica para assuntos de saúde como fonte de termos do consumidor. Neste estudo a aplicação de técnicas de reconhecimento automático de termos e validação manual resultou na obtenção de cerca de 250 termos. Jiang et al.[31] propuseram a extração de termos utilizados em textos de redes sociais por meio da aplicação de técnicas estatísticas, exclusivamente para termos relativos a reação adversa de drogas.

Kim et al.[32] reportaram a construção de um CHV baseado na extração de termos provenientes de artigos disponíveis na web. Vydiswaran et al.[23] reportaram a obtenção de termos do consumidor e técnicos por meio da recuperação de artigos da Wikipedia. Chen et al.[33] utilizaram o conteúdo da OAC-CHV como base para o desenvolvimento de um sistema baseado em processamento de linguagem natural que identifica automaticamente e classifica termos usados por médicos em registros eletrônicos de saúde. He et al.[34] analisaram conteúdos disponibilizados por usuários em portais web, essencialmente perguntas e respostas, e mostraram que é uma fonte promissora de termos utilizado por consumidores em saúde.

Uma aplicação prática baseada no conteúdo da OAC-CHV foi descrita por Zeng-Treitler et al.[35] Um protótipo foi desenvolvido para realizar a tradução automática de termos técnicos para termos leigos.

Em relação à utilização de tecnologias da web semântica, o estudo de Tilahun et al.[29] apresenta uma proposta de desenvolvimento de um aplicação baseada em dados abertos, conectando bases como DBpedia, Bio2RDF e LinkedCT, disponíveis em formato de interconexão de dados RDF. Bushman et al.[36] descreveram o projeto MeSH RDF, que pretende facilitar o compartilhamento e a vinculação de dados usando padrões e tecnologias semânticas da web.

1.4 Organização da tese

Esta tese está organizada da seguinte forma:

Capítulo 1 Introdução: o capítulo corrente aborda os elementos que subsidiaram os objetivos gerais e específicos desta tese, no qual foram explorados tópicos referentes aos estudos sobre o desenvolvimento de CHVs, princípios da web semântica, assim como os trabalhos correlatos a esse estudo.

Capítulo 2 Fundamentos Teóricos: são apresentadas ao leitor explicações dos conceitos que suportam essa tese;

Capítulo 3 Objetivo: são apresentados ao leitor os objetivos gerais e específicos;

Capítulo 4 Método: são detalhados os processos e aplicação das técnicas que proporcionaram alcançar cada objetivo específico;

Capítulo 5 Resultado: são apresentados os resultados dos experimentos realizados;

Capítulo 6 Discussão: discussão e comparação dos métodos aplicados e resultados obtidos em relação a outros estudos;

Capítulo 7 Conclusão: síntese sobre a contribuição do estudo e possibilidades para sua continuidade.

2 OBJETIVO

O objetivo desta pesquisa foi desenvolver um vocabulário de saúde para consumidor (VSC) em português composto por conteúdos disponíveis na web e estruturado segundo conceitos e tecnologias da web semântica.

Os objetivos específicos constam como:

1) Construir uma rede complexa para estruturar um conjunto de termos técnicos e termos de domínio do consumidor e analisar as relações estabelecidas entre eles.
2) Identificar termos e relações semânticas ainda não mapeados no Objetivo 1 e validá-los por meio de avaliação humana.
3) Construir uma aplicação computacional para acesso ao conteúdo do VSC.

2.1 Justificativa

A questão a ser explorada nesse estudo é como podem ser construídos o conteúdo e a estrutura de um vocabulário de saúde do consumidor (VSC) em português que possibilite suportar aplicações para auxiliar o consumidor de saúde em suas buscas por informação em saúde de forma que ele utilize termos de seu próprio domínio. A tese essencial aqui apresentada é que por meio da análise de conteúdos disponíveis na web é possível extrair termos usados por consumidores e associá-los aos termos presentes em vocabulários controlados. A partir dessa extração é possível construir um vocabulário de forma que relacionamentos, por exemplo, sinonímia, entre seus elementos estejam identificados, fator importante para a recuperação da informação e para sua utilização na leitura por máquinas.

Este projeto compõe a iniciativa Saúde 360º (saude360.unifesp.br) que visa obter a melhor experiência do consumidor quanto à busca por informação de saúde na web. Nossas pesquisas visam conectar diferentes bases de dados que contenham algum conteúdo de saúde, como redes sociais, listas de medicamentos, vocabulários controlados, cadastros de estabelecimentos de saúde, portais de notícias entre outros, e evidenciar informações relacionadas nessas bases.

3 FUNDAMENTOS TEÓRICOS

Neste capítulo são apresentados os conceitos e definições que suportam o desenvolvimento deste estudo.

As definições que suportam esse estudo, relacionadas à terminologia linguística adotada foram sistematizadas em um glossário (Glossário, p. 127).

3.1 Definições básicas

Este estudo segue as definições adotadas pelo glossário do UMLS segundo o qual termo é definido como “uma palavra ou coleção de palavras que compreende uma expressão”, assim como “um termo é a classe de todas as cadeias que são variantes uma da outra, por exemplo, “olho” (singular) e “olhos” (plural) são termos que englobam um mesmo conceito”.[37] O UMLS agrupa termos com um significado exclusivo em um conceito com identificador único, o Concept Unique Identifier (CUI). Todos os termos dentro de um conceito são sinônimos.[37]

O UMLS é um sistema de terminologias [38] composto por Metathesaurus, Semantic Network, SPECIALIST Lexicon e Lexical Tools Sources Map.[39] Metathesaurus é o maior componente, composto por vocabulários que registram conceitos, termos e relações (sinonímia, hiponímia, hiperonímia) entre eles estabelecidos entre sistemas como CID-10, SNOMED e o Medical Subjects Headings (MeSH). Cada conceito no UMLS é descrito por um termo preferido e seus sinônimos. A rede semântica fornece informações sobre o conjunto de categorias semânticas básicas, os tipos semânticos, a saber, organismos, doença ou síndrome, processos, entre outros.[38]

Para a edificação deste estudo uma opção conveniente foi herdar as denominações utilizadas na OAC-CHV, essencialmente baseadas nos vocabulários do UMLS. Nestes, a relação de sinonímia entre termo preferido e outros termos é estabelecida por meio do CUI, que usado para unir a OAC-CHV e outros vocabulários do UMLS, como o MeSH e o Medical Dictionary for Regulatory Activities Terminology (MedDRA).

Os aspectos contextuais referentes aos possíveis significados do CUI foram

descritos por Campbell et al.[40], que argumentam que o UMLS contém proposições de forma que um conceito possa ser associado a um conjunto de termos. Proposições são representações de palavras, pensamentos e objetos, que são os elementos básicos do triângulo semântico de Ogden-Richards.[41] O triângulo de Ogden-Richards foi desenvolvido para explicar a relação da linguagem (palavras) com os pensamentos e com o mundo (objeto). Campbell et al.[40] argumentam que o CUI une as representações de um mesmo conceito por extensão (denotação), como definições, categorizações e os sinônimos. Entretanto, a intensão (conotação) do conceito pode ser diferente nos diversos vocabulários. O mesmo conceito pode ter significados diferentes no MeSH e SNOMED, porém podem ser identificados pela categoria associada ao conceito durante o desenvolvimento desses itens. Outro aspecto considerado é que o significado do conceito evolui à medida que novos termos são incluídos ou excluídos.

O presente estudo visa mapear as relações estabelecidas entre o signo (termo) e o objeto, ou seja, no nível semântico. A semântica dedica-se a identificar o significado que uma palavra ou um texto assumem quando inseridos em determinado contexto, o que é importante para sistemas de representação na área de saúde porque esta faz uso de termos comuns cujo significado é diferente em outras situações.

Desta forma, a definição de Obrst [42], que considera termos como palavras ou frases em linguagem natural, usada para indicar a semântica (significado), isto é, o conceito, cabe neste estudo porque aqui a relação entre termo e conceito é estabelecida, o que recupera a ideia de conceito do UMLS abordada neste trabalho.

A definição de conceito, à luz de ontologias na visão das ciências da computação, também é referida por Obrst [42] e torna-se importante neste estudo porque considera conceito como uma unidade semântica, uma entidade ou ligação em modelos de representação do conhecimento. Em uma ontologia um conceito é uma unidade de conhecimento primária que descreve classes, relações, atributos e indivíduos (objetos básicos), associados a um domínio, que são interpretáveis pela máquina.

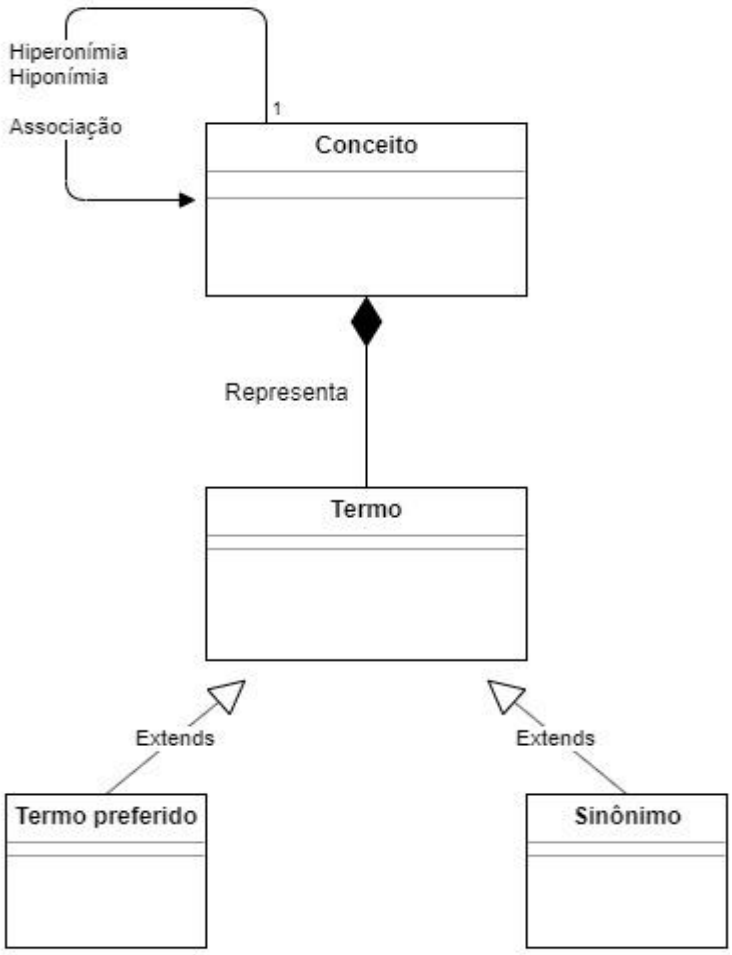
Resumidamente, para este estudo os termos são usados para representar o conceito como uma classe de objetos que podem ser considerados equivalentes.[40]

Os conceitos são descritos por meio do termo preferido e estendido por meio de outros termos (sinônimos), definição e categorização. Por herança do UMLS os conceitos são representados por meio de um identificador, o CUI, que une os termos por extensão, o que possibilita a inserção de novos termos de forma a manter o significado do mesmo no conjunto, o que é um elemento importante na construção desse estudo.

O mapeamento das categorias, que associam os conceitos a outros conceitos com significados ampliados por meio de relações hierárquicas de hiperonímia, também é importante para este estudo. As relações hierárquicas não contemplam apenas hiperonímia/hiponímia e, em alguns casos, podem conter relacionamentos entre classe/subclasse ou parte/todo, de acordo com a descrição do MeSH.[43] Relacionamentos entre conceitos também podem ser estabelecidos por meio de associação por ocorrência conjunta, a exemplo de “bactéria” e “antibiótico”.

A Figura 1 mostra a visão de conceito-termo e suas relações semânticas que suportam este estudo. Entende-se que termo é referente à representação do conceito e por isso é composto por termo preferido e seus sinônimos. Para um mesmo conceito a relação entre termos é de sinonímia. Entre os conceitos, representados por termos preferidos, as relações podem ser estabelecidas entre o termo e sua categoria (hiperonímia) ou por associação.

Alguns estudos mostram que os possíveis relacionamentos entre termos, ou mesmo palavras, podem ser formalizados e analisados por meio de redes complexas, que podem estar arranjados em textos [44] ou mesmo em um tesauro.[45] Outro aspecto importante para esse estudo é que as redes complexas podem ser divididas em sub-redes, construídas de acordo com algumas características a serem determinadas, mas que podem auxiliar na verificação da consistência dos termos e suas associações, assim como na exploração de conjuntos de termos que não estão associados por extensão a um conceito.



Fonte: Modificada de Vandenbussche e Charlet.[46]

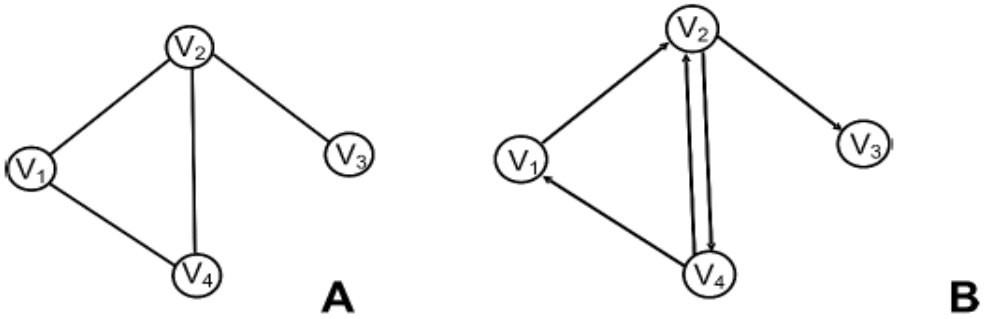
Figura 1. Representação UML referente à visão conceito-termo e suas relações adotadas neste estudo. O conceito é representado por um conjunto de termos relacionados por sinonímia.

3.2 Redes complexas

Redes complexas consistem de grafos e de informações adicionais sobre seus elementos estruturais.[47] Nos campos da biologia e da medicina as aplicações constam de identificação de alvos de drogas, determinação da função de uma proteína ou gene, planejamento de estratégias efetivas para tratar doenças ou diagnóstico precoce de distúrbio.[48] São importantes para conteúdos biomédicos, que podem ser analisados em função de suas relações e hierarquias, elementos inerentes às redes.

Um grafo G = (V,E) é formado por um conjunto V de pontos ou nós,

denominados vértices, e de um conjunto E de pares não ordenados de linhas que os conectam, as arestas. Quando as linhas são direcionadas os arcos formam um grafo direcionado.[49] Essa estrutura possibilita a representação de dados por meio dos vértices e de suas relações por meio das arestas (Figura 2).



Fonte: Extraído parcialmente de Pavlopoulos et al.[48]

Figura 2. Grafos não direcionado (A) e direcionado (B). V_1 a V_4 indicam os vértices e os segmentos de reta, as arestas.

Citraro e Rossetti [50] afirmam que as redes semânticas complexas, que são estruturas para representação do conhecimento, são grafos G = (V, E) em que V são os termos ou palavras conectados por algum tipo de propriedade semântica E, determinada a partir do aspecto que elas estão modelando. As propriedades semânticas representam relações como sinonímia, hiponímia, hiperonímia, associações livres ou os argumentos de verbos.

Dessa forma um vocabulário pode ser suportado por um grafo - um grafo direcionado G = (V, E): os termos determinam o conjunto de vértices V; e $E \subseteq V \times V$ é o relacionamento entre vértices. Esse modelo de vocabulário possibilita buscas por termos importantes (centrais) nas redes, partes importantes e visualização do conjunto de termos que o compõe.[51]

Alguns conceitos sobre grafos a serem abordados nesta tese estão relacionados a seguir. Critérios de classificação podem ser aplicados de acordo com medidas relacionadas aos vértices. As medidas de centralidade são indicativas da importância do vértice na rede.

As medidas utilizadas nesta tese foram:

- Grau: número de arestas ou arcos relacionados ao vértice. Consideram-se graus de entrada e de saída no caso de grafos

direcionados.[49] O grau indica o número de interações entre os vértices.[48] Formalmente, considerando G = (V,E) um grafo, V o vértice e E a aresta , o grau $d_G\ (v)$ de um vértice *v* é o número |E(*v*)| de arestas em *v*, que é igual ao número de vizinhos de *v*. Para o cálculo de centralidade C de grau d considera-se que $C_d\ (v) = d(v)$.[48]

- Proximidade (*closeness*): é o número de outros vértices dividido pela soma de todas as distâncias geodésicas entre o vértice e todos os outros. Para uma rede direcionada, essa medida é denominada *proximity prestige*; é calculada em função do grau de entrada. Essa medida indica os nós importantes que podem se comunicar rapidamente com outros nós da rede.[48] Formalmente a centralidade *C* de *closeness* é calculada pela equação 1:

$$C_{\text{clos}}(v) = \frac{1}{\sum_{i\,\epsilon\,V}^{|V|} dist(i,j)} \qquad (1)$$

onde *dist(i,j)* é a distância geodésica ou o menor caminho entre os vértices *i* e *j*.[48]

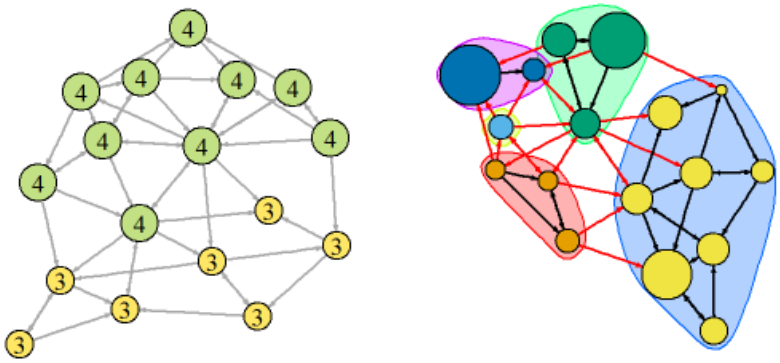
- Intermediação (*betweeness*): é a proporção de todas as distâncias geodésicas entre pares de outros vértices que incluem este vértice. Essa medida atribui maior valor aos nós que são intermediários entre os vizinhos.[48] Formalmente, considerando os vértices *i,j,w* ∈ V(G), seja σ_{ij} o número total de caminhos mínimos de *i* a *j* e $\sigma_{ij}(w)$ o número total de caminhos mínimos de *i* a *j* que passam por *w*. A centralidade de *betweeness* é calculada pela equação 2:

$$C_b(w) = \sum_{(i\ j)\,\in\,V(w)} \frac{\sigma_{ij(w)}}{\sigma_{ij}} \qquad (2)$$

Grafos também podem ser agrupados para fins de análise de características comuns ou para melhor visualização de características específicas. O objetivo do agrupamento (cluster) é organizar objetos diferentes observando propriedades comuns de elementos em um sistema.[48]

Agrupamentos podem ser obtidos aplicando algoritmos específicos baseados em:

- Núcleos (*core*): o grau pode ser utilizado para identificar agrupamentos com o mesmo grau e compor uma sub-rede (Figura 3a).[47]
- Detecção de comunidades: identificação de subgrafos (geralmente chamados de comunidades, clusters ou módulos) que compartilham propriedades comuns e/ou desempenham papéis semelhantes no grafo. O objetivo é identificar os módulos e, possivelmente, sua organização hierárquica, usando apenas as informações codificadas na topologia do grafo (Figura 3b).[52]



Fonte: extraído de Ognyanova.[53]

Figura 3. Agrupamentos de vértices baseados em núcleos ou *core* (a) e comunidades (b).

3.3 Análise de textos

A aplicação de técnicas computacionais com o objetivo de extrair, recuperar e compreender elementos constituintes de conteúdos não estruturados compõe o campo da ciência da computação denominado mineração de texto.[54] Atualmente é denominada análise de texto (*text analysis*). Trata-se da aplicação de técnicas para transformar textos em dados estruturados que possibilitam uma análise global para fins de extração ou recuperação automatizada da informação. Para a área de saúde torna-se útil para o processamento automatizado da literatura e desencadeou

pesquisas referentes a problemas de identificação de termos em textos biomédicos.

A recuperação é a primeira etapa do processo identificação de termos, seguida por classificação, que associa o termo a uma categoria, e mapeamento, que associa o termo a vocabulários controlados ou bases de dados.[55]

A aplicação de técnicas baseadas em análise linguística e estatística é a base para a tarefa de recuperação de termos.[56] Especificamente, a aplicação dessas técnicas também possibilitam a extração de palavra-chave (*keyword*) ou frase-chave (*keyphrase*) do texto.[57, 58]

As seções seguintes detalham a aplicação dessas técnicas.

3.3.1 Técnicas avançadas de reconhecimento automático de termos

Reconhecimento Automático de Termos (RAT) é o processo que executa a extração de termos pertencentes a um determinado domínio por meio do uso de computadores.[56] RAT provê a habilidade de manipular um grande conjunto de dados, uma impossibilidade para os processos manuais. RAT é especialmente recomendado para a construção de vocabulários especializados.[56]

RAT é a principal etapa para encontrar termos importantes para a construção de um vocabulário. Nela, são utilizadas técnicas baseadas na estrutura linguística e estatística de textos a fim de reconhecer termos importantes ou que se destacam em um texto.

As técnicas de RAT apresentam duas abordagens: linguística, que envolve a aplicação de algoritmos que classificam o termo de acordo com regras morfossintáticas pré-definidas, e estatística, que é baseada na hipótese de que a importância do termo está relacionada à sua frequência de uso no domínio.[59]

Para extrair um conjunto de termos candidatos a abordagem linguística possibilita a identificação de classes gramaticais (*part-of-speech - POS),* sintagmas nominais (*noun phrases*) e entidades nomeadas (*named entities recognition - NER*).

A abordagem estatística avalia o peso de cada termo candidato [58], o que pode ser obtido por meio da frequência relativa do termo em relação aos documentos, calculado por meio do tf-idf. O tf-idf determina a frequência relativa de palavras em um determinado documento em comparação com a proporção inversa

dessa palavra sobre todo o corpus do documento. Intuitivamente esse cálculo determina a relevância de uma determinada palavra em um documento específico.[60]

Outra abordagem possível é um método híbrido, que combina marcação linguística e análise estatística de forma a aproveitar todas as vantagens que cada uma oferece, com o objetivo de aplicar as técnicas de RAT de forma mais eficiente. Uma das aplicações de abordagem híbrida é o c-value,[61] muito utilizado em estudos cujo objetivo é obter termos pertencentes ao domínio biomédico [30], assim como em outros domínios e idiomas diferentes do inglês.[62] Nessa abordagem a marcação linguística constrói a lista dos termos candidatos, enquanto a parte estatística designa a medida de termo para cada termo na lista, ou seja, classifica os termos na lista.

A parte linguística consiste em: [61]

1. Marcação linguística: classificação dos termos segundo classe gramatical (etiquetagem POS), como “substantivo”, “verbo”, “adjetivo” etc.
2. Filtro linguístico: aplicado ao corpus marcado para excluir palavras cujas classes gramaticais não são necessárias para extração.
3. *Stopwords*: exclusão de palavras que não poderiam ser classificadas como termos, como conjunções e palavras muito frequentes.

A análise estatística classifica o termo candidato cujo cálculo baseia-se em:[57]

1. Peso: cálculo baseado em contagem, frequência ou tf-idf.
2. Localização do termo no documento.

Um dos problemas dessas abordagens é estabelecer um limite arbitrário para considerar o conjunto de termos válidos. Recentemente algoritmos de aprendizado de máquina têm sido aplicados para estabelecer um conjunto limitado de termos. Esses algoritmos apresentam vantagens como a capacidade de aprender a reconhecer um termo e a facilidade do uso de um grande número de medidas, de forma a reduzir o tempo para execução dos processos.[63]

A aprendizagem de máquina apresenta melhor desempenho quando aplicada

no reconhecimento de palavra/frases-chave, mas também pode ser aplicada em reconhecimento de termos.[64] Há duas abordagens possíveis para o processo de aprendizagem: supervisionado, na qual o algoritmo é executado segundo um conjunto de treinamento e cujos resultados dependem da qualidade do vocabulário relacionado ao documento, e não supervisionado, que depende unicamente da estrutura do documento e das relações estatísticas entre os termos no documento.

Diversos algoritmos foram desenvolvidos segundo essas abordagens para fins de extração automática de palavras/frases-chave[57], por exemplo, o Rapid Automatic Keyword Extraction (RAKE). Este é um algoritmo não supervisionado, independente de domínio e de idioma desenvolvido para extrair palavras-chave de documentos individuais.[65] Segundo Rose et al.[65] uma palavra-chave é uma sequência de uma ou mais palavras que fornece, idealmente, o conteúdo essencial do documento.

3.3.2 Rapid Automatic Keyword Extraction (RAKE)

RAKE baseia-se na observação de que as palavras-chave geralmente contêm várias palavras, mas elas raramente contêm pontuação ou *stopwords*.[65] Os parâmetros de entrada para RAKE incluem uma lista de *stopwords*, um conjunto de delimitadores de frase e outro de palavras, que são utilizados para particionar o texto em palavras-chave. Coocorrências entre palavras são significativas e são usadas para marcar as palavras-chave candidatas.

Após a identificação de cada palavra-chave candidata, as coocorrências são mapeadas. Uma pontuação é calculada para cada palavra-chave candidata. A métrica para calcular as pontuações de palavras é baseada no grau (número de interações) e na frequência da palavra vértices no grafo construído automaticamente pelo algoritmo: (1) frequência de palavras (freq (w)), (2) grau de palavra (deg (w)) e (3) relação de grau para frequência (deg (w) / freq. (w)).

Após a pontuação das palavras-chave candidatas, o algoritmo apresenta como resultado, por padrão, um terço do total de palavras melhor pontuadas.

RAKE apresentou melhores valores de precisão e *F-measure* que outros métodos de aprendizagem supervisionada na tarefa de extrair palavras-chave a

partir de resumos.[65] Outro aspecto importante é que RAKE mostrou-se mais eficiente, em relação ao tempo de processamento, que outros algoritmos.

4 MATERIAIS E MÉTODOS

Esta seção descreve as ferramentas e processos aplicados na construção do VSC. Inicialmente, é apresentado o processo para a elaboração de uma revisão da literatura para fins de verificação do desenvolvimento atual da OAC-CHV ou outros modelos. Em seguida, apresenta-se a descrição do desenvolvimento do VSC, realizado em três fases. A Figura 4 mostra a sequência de execução das fases metodológicas e a relação com os objetivos propostos.

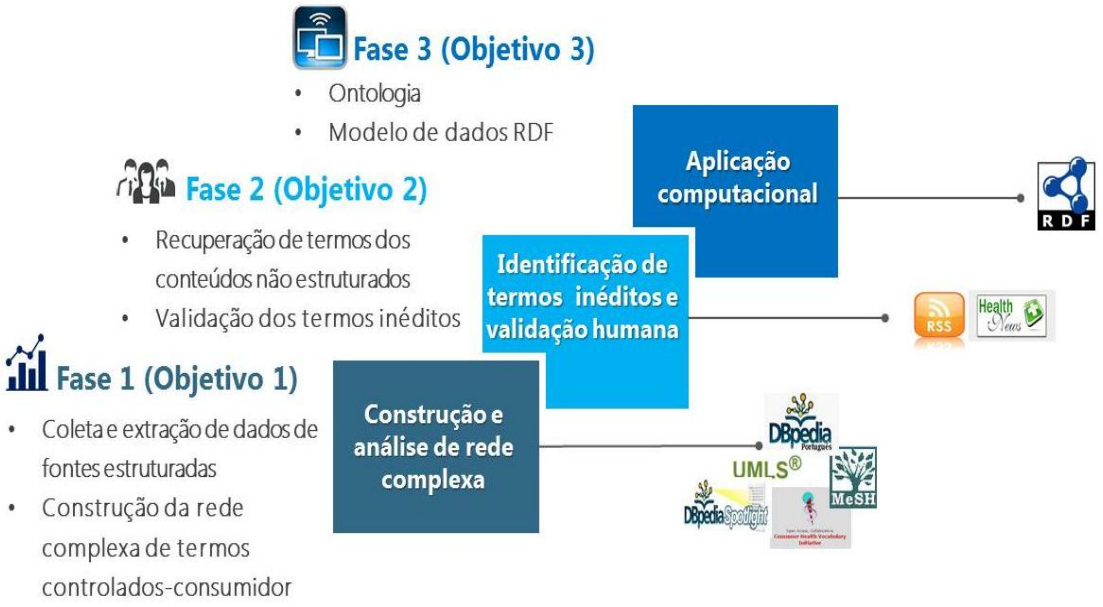


Figura 4. Fases metodológicas associadas aos objetivos específicos.

A Fase 1 foi referente à representação dos termos candidatos e suas relações semânticas por meio de uma rede complexa, contemplando o Objetivo 1. Para a composição do conteúdo foram coletadas e armazenadas localmente bases de dados compostas por termos, codificações e categorias, disponíveis para download. Além disso foram executados processos automatizados para extração de dados que constam de identificadores, termos e relações semânticas, disponibilizados por meio de *webservices*. Os termos e relações semânticas foram representados por meio de uma rede complexa, o que possibilitou realizar uma análise para obtenção de características referentes ao relacionamento entre estes.

O Objetivo 2 foi contemplado na Fase 2, que consta das ações para

identificação de termos de saúde inéditos provenientes de conteúdos não estruturados. Para isso, foi descrita a aplicação de técnicas de análise de texto, além das estratégias usadas para a validação humana dos termos candidatos e das relações estabelecidas entre eles.

Finalmente, de acordo com o Objetivo 3, a Fase 3 descreveu o processo para obter a representação do conteúdo do VSC baseada em um modelo de interconexão de dados RDF e o desenvolvimento de uma aplicação para explorar o conteúdo.

4.1 Tipo de estudo

Os objetivos específicos são suportados pelo tipo de pesquisa exploratória, de caráter descritivo e abordagem quantitativa.[66]

Esse estudo tem por base a linha de pesquisa intitulada Registro, Recuperação e Relacionamento de Dados em Saúde junto ao Programa de Pós-graduação em Gestão e Informática em Saúde EPM/UNIFESP.

4.2 Comitê de Ética em Pesquisa e conflito de interesses

Este estudo foi aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de São Paulo (#125.943/12), em 05/10/2012 (Apêndice 1 – Aprovação do Comitê de Ética em Pesquisa, p. 117).

Os autores declararam não haver conflito de interesse na condução dessa pesquisa. Esta pesquisa não foi conduzida com a participação de pacientes diretamente, não oferecendo fatores de risco a eles.

Esse estudo contou com a participação de especialistas voluntários da área de saúde e informática em saúde que foram convidados por meio de divulgação em mídia eletrônica, durante a Fase 2, para a realização de avaliação que visava a classificação de termos candidatos ao VSC. Não houve pagamento financeiro aos voluntários.

4.3 Revisão da literatura

No período entre 18 e 25 de junho de 2019 foram realizadas buscas nas bases PubMed (www.ncbi.nlm.nih.gov/pubmed), Portal de Periódicos Capes/MEC (www.periodicos.capes.gov.br/) e Google Scholar (scholar.google.com) com a finalidade de atualizar uma revisão realizada quando no início do projeto sobre estudos que relatam a aplicação de métodos para atualização ou aplicações baseadas na OAC-CHV. Além disso, buscou-se identificar técnicas para o desenvolvimento de outros modelos de CHV.

De forma a especificar o escopo da estratégia de pesquisa utilizou-se a expressão *“consumer health vocabulary” OR "lay term" OR "consumer health expressions" OR "consumer health terminology"* em cada uma das bases relacionadas.

Os artigos originais publicados no período entre 2014 e 2019 (incompleto) em idioma inglês atendem aos critérios de inclusão neste estudo. As exclusões constam de artigos que não foram submetidos à avaliação por pares ou com resumos indisponíveis e duplicações entre as bases pesquisadas.

De forma sequencial títulos, resumos e textos completos foram lidos e os artigos que não descreviam o desenvolvimento ou aplicações de CHV foram excluídos em cada refinamento.

4.4 Fase 1 - Construção de rede complexa de termos (Objetivo 1)

Nesta seção é descrita a construção das bases de dados provenientes de fontes estruturadas e representação dos termos por meio de redes complexas. Os dados são referentes a termos, identificadores e relações semânticas como hiperônimos, sinônimos e outras.

As bases de dados estruturadas constam dos vocabulários controlados do UMLS, disponíveis em idioma português, associados ao conteúdo da OAC-CHV e dos Descritores em Ciência da Saúde (DeCS), e uma base de conhecimento de domínio aberto criada a partir de conteúdo colaborativo, a DBpedia (pt.dbpedia.org).

Neste estudo os termos provenientes dos vocabulários UMLS/OAC-

CHV/DeCS foram denominados “termos controlados” e os termos da DBpedia, “termos do consumidor”, porque são provenientes de conteúdos textuais escritos por usuários da Wikipedia. Os conteúdos de saúde e medicina foram, no contexto desse estudo, escritos por e para consumidores em saúde. Os termos do consumidor são denominados termos leigos (*lay terms*) em estudos semelhantes.[9, 16]

A Figura 5 apresenta o fluxo para a coleta e extração de dados de bases estruturadas, assim como a construção da rede complexa de termos controlados e do consumidor. As seções seguintes apresentam os detalhes referentes a cada item.

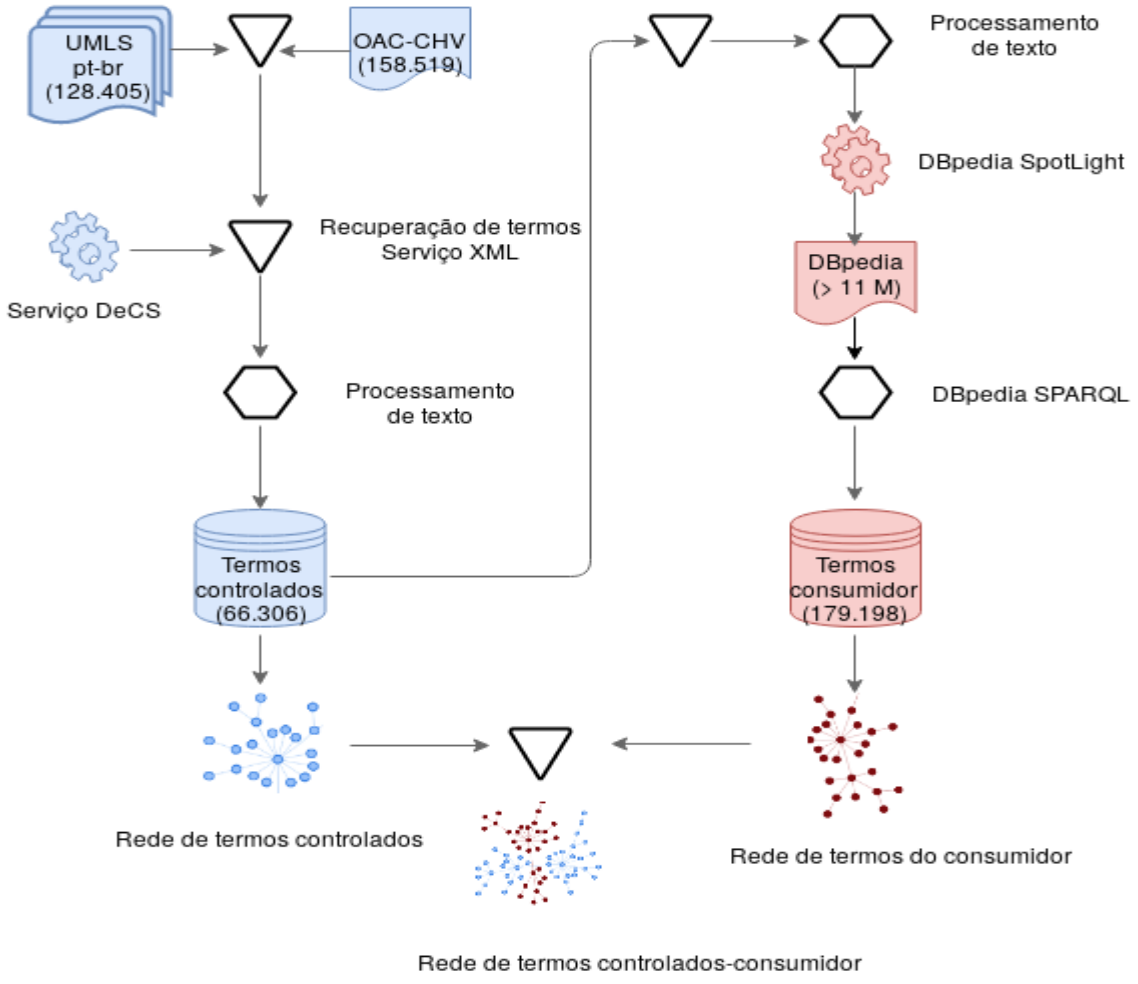


Figura 5. Fluxo para coleta e extração de dados e construção da rede complexa de termos controlados-consumidor.

4.4.1 Construção das bases de dados

Uma breve descrição sobre as bases utilizadas e ferramentas para extração de dados é apresentada nas seções seguintes.

4.4.2 Descrição das bases de dados

A coleta de dados dos vocabulários controlados UMLS® Metathesaurus® disponíveis na versão português foi intermediada pela ferramenta MetamorphoSys. A versão utilizada foi 2015AB Metathesaurus. Para o idioma português são disponibilizados quatro dicionários controlados: Medical Dictionary for Regulatory Activities Terminology (MedRA) (95.500 termos - 50.041 conceitos únicos), Medical Subjects Headings (MeSH) (65.150 termos - 34.470 conceitos únicos), International Classification of Primary Care (ICPC) (725 termos) e Terminology for Coding Clinical Information in Relation to Drug Therapy (WHO) (3.750 termos). Os quatro vocabulários contêm 128.405 termos únicos e 79.790 conceitos únicos.

A OAC-CHV provê o acesso ao arquivo composto por 158.519 termos, sendo 57.819 conceitos únicos associados a um identificador único, o CUI, que é comum às bases.

Adicionalmente utilizamos os dados extraídos do DeCS, um vocabulário trilíngue (português, espanhol e inglês), construído a partir do MeSH, que inclui temáticas de saúde da América Latina e Caribe. Possui 33.136 termos preferidos, denominados descritores, sendo destes 28.552 provenientes do MeSH e 4.584 exclusivos.[67]

Além das categorias mapeadas pelo MeSH (Doenças, Anatomia, Compostos Químicos e Drogas etc.), o DeCS também disponibiliza a categorização dos termos como Saúde Pública, Ciência e Saúde, Homeopatia e Vigilância Sanitária.[67] Um serviço para acesso ao conteúdo do DeCS está disponível na web (decs.bvsalud.org/cgi-bin/mx/cgi=@vmx/decs/?words=).

A DBpedia é um projeto colaborativo que visa extrair conteúdo estruturado da Wikipedia e disponibilizá-lo por meio das tecnologias da web semântica.[19, 68] Uma das tecnologias que suportam a DBpedia é a ferramenta DBpedia Spotlight

(dbpedia-spotlight.org) que recupera automaticamente anotações dos recursos da DBpedia em textos.[69]

4.4.3 Coleta e extração de dados

Os dados dos vocabulários controlados do UMLS foram coletados do arquivo MRCONSO.RRF, que contém identificadores de conceito como o CUI, o Distinct Unique Identifier (DUI) e os termos por extenso.[37]

A OAC-CHV propiciou a coleta de um arquivo composto pelos campos CUI e o termo por extenso, em idioma inglês. Verificou-se que o CUI é uma chave única entre os vocabulários do UMLS coletados e a OAC-CHV e foi utilizado para filtrar os dados. O resultado foi um subconjunto de termos do UMLS identificados previamente como de domínio do consumidor, em idioma português.

Esses termos foram utilizados para extrair dados em formato XML disponibilizados pelo serviço DeCS. Para cada termo foram recuperados a categoria, sinônimos e outros termos preferidos, representativos dos conceitos, associados por hierarquia e por coocorrência. Este processo possibilitou constituir uma base de dados composta por identificadores, termos e relações semânticas, provenientes dos vocabulários controlados. Para simplificar, esta base será referida como base de dados controlados.

Os termos da base de dados controlados foram testados para verificar a ocorrência dos termos no domínio do consumidor em saúde. Um serviço RESTful DBpedia SpotLight hospedado em um servidor no Datacenter de Pesquisa da UNIFESP, localizado no Departamento de Informática em Saúde, suportou a busca pelos termos. Este serviço foi utilizado para anotação automática dos termos mencionados nos recursos DBpedia. Para cada termo anotado foi recuperado o Uniform Resource Identifier (URI). Os parâmetros utilizados no DBpedia SportLight foram *confidence*=0,2 e *support*=20. Outros valores foram testados, porém os valores adotados conferiram maior flexibilidade na recuperação do termo exato (considerando as formas plural e singular) ou parte do termo.

Um script foi desenvolvido para acessar a URI referente a cada termo e recuperar os valores referentes a propriedades disponíveis na versão da DBpedia

em idioma português (pt.dbpedia.org) usando DBpedia SPARQL.[19]

As propriedades da DBpedia recuperadas foram análogas à proposta para a base de dados controlados. O Quadro 1 mostra as propriedades recuperadas e sua relação com a estrutura construída para a base de termos controlados. A propriedade *dbpedia-owl:wikiPageRedirects* não refere o sentido estrito dos sinônimos. Liu[70] considera que cerca de 78% dos *redirects* são sinônimos. O restante é referente aos redirecionadores, ou seja, um termo com significado estendido mas condizente com o conceito, o que é importante para a navegação e está de acordo com as potencialidades do VSC.

Quadro 1. Propriedades recuperadas da DBpedia relacionadas à estrutura da base de termos controlados.

DeCS*	DBpedia	Comentário
<term list>	rdf:type	Categoria
<synonim>	dbpedia-owl:wikiPageRedirects	Sinônimo estendido
<related>	dbpedia-owl:wikiPageWikiLink	Coocorrência
<descriptor>	rdfs:label	Termo preferido

* Tags XML recuperadas.

Os valores referentes à propriedade *dcterms:subject* também foram recuperados por associar o termo a outros com significado mais abrangente, por exemplo, “insulina” com “hormônio pancreático” e “endocrinologia”. Trata-se de um mapeamento importante neste estudo para estabelecer relações semânticas do termo com uma categoria.

Após a recuperação do conteúdo foi executado um pré-processamento de texto para fins de remoção de acentuação e pontuação, de forma a unificar os termos com variações de escrita. Além disso, foram excluídos termos referentes a notações da DBpedia, como indicação de artigos excluídos ou listas. Um ajuste necessário foi estruturar as classes referentes a ontologia DBpedia (mappings.dbpedia.org/server/ontology/classes). Por exemplo, havia termos categorizados como "músculo", mas não como "estrutura anatômica". Ao importar para a base de dados a relação entre essas categorias foi possível estabelecer uma ligação entre elas que se estendeu a todos os termos recuperados e, ainda, casos com dupla categorização. Outro ajuste realizado foi a exclusão arbitrária (manual) de siglas e acrônimos com significados diferentes dos vocabulários UMLS/DeCS. Essa

inspeção foi realizada pela própria pesquisadora por meio da observação dos termos em listagem.

Esse processo resultou na base de dados do consumidor.

4.4.4 Construção da rede complexa

Os termos obtidos das fontes de dados estruturados foram organizados para fins da construção de duas redes complexas direcionadas, identificadas neste estudo como redes de termos controlados e do consumidor. Nesta tese os vértices representam os termos e as arestas representam as relações semânticas. A estruturação em rede direcionada será considerada na Fase 3.

As redes de termos controlados e do consumidor foram construídas por meio do software Pajek versão 4.07 64 bits (mrvar.fdv.uni-lj.si/pajek/).[47, 71] Pajek é especialmente projetado para lidar com grandes conjuntos de dados por disponibilizar ferramentas de visualização poderosas, além de implementar uma variedade de algoritmos eficientes para o cálculo de métricas que suportam a análise da rede.[47]

Pajek oferece recursos para a união de redes, o que foi aplicado para unir as redes de termos controlados e do consumidor por meio do vértice. Os termos comuns foram unificados automaticamente, o que resultou em uma rede de termos controlados-consumidor.

Algoritmos foram executados para obter medidas de centralidade para cada rede construída. As medidas possibilitaram identificar os termos importantes, cujo significado pode revelar características próprias das relações entre esses. As medidas de centralidade obtidas foram *proximity prestige*, um valor indicativo da influência ou prestígio do vértice, *betweenness*, que indica a importância do vértice para a conexão de outros termos, além do grau, cujo valor indica o número de ligações do vértice.[47, 48]

Devido à grande quantidade e diversidade de assuntos referentes aos termos recuperados por meio da DBpedia foi necessário estabelecer um processo semiautomático que possibilitasse selecionar termos caracterizados como de domínio da área de saúde, consistentes com o modelo de vocabulário em

desenvolvimento. Para minimizar a necessidade de estabelecer critérios empíricos para seleção foi aplicada uma abordagem não supervisionada, baseada na execução de um algoritmo para identificar agrupamentos da rede complexa. Desta forma os agrupamentos são identificados usando apenas as informações codificadas na topologia da rede complexa.

Algoritmos para identificar e visualizar agrupamentos foram executados por meio do software Pajek. Este processo é detalhado na próxima seção.

4.4.5 Seleção de termos válidos para o VSC

Agrupamento ou clusterização é um processo que envolve a separação de dados em conjuntos similares. Para essa tarefa os métodos tradicionais aplicados para o reconhecimentos desses agrupamentos são baseados em particionamento e clusterização hierárquica por divisão ou aglomeração.[72] Clusterização hierárquica é utilizada para a detecção de comunidades em uma rede complexa.

O objetivo da detecção de comunidades em uma rede complexa é identificar módulos e limites que possibilitam a classificação de vértices de acordo com a topologia do grafo.[52] Uma comunidade é um subgrafo cuja coesão entre os vértices do módulo é muito maior que com o restante do grafo.[52] Assume-se que os vértices que compõem uma comunidade apresentem mais similaridade entre si de acordo com alguma medida.[73]

Louvain [74] é um dos algoritmos de detecção de comunidades baseado em clusterização hierárquica aglomerativa, que utiliza como medida a maximização da modularização.[73] O cálculo da modularização considera que um subgrafo é uma comunidade se o número de arestas entre seus vértices excede o número esperado de arestas que o mesmo grafo poderia ter de forma aleatória.[74]

Identificar um conjunto de vértices similares é uma tarefa importante neste estudo. Trata-se de separar termos ligados entre si que apresentem características próximas dos termos coletados dos vocabulários do UMLS. Outro aspecto importante é que a detecção de comunidades é baseada em abordagem não supervisionada, na qual grupos densos de vértices com características estruturais similares podem ser encontrados [75] sem depender de escolhas empíricas.

O software Pajek disponibiliza o algoritmo Louvain e permite a variação de alguns parâmetros como *resolution* (resolução), que controla o número de comunidades encontradas, máximo número de níveis e repetições em cada iteração. Há duas possibilidades de escolha do algoritmo em relação ao refinamento das comunidades: é executado apenas no último nível ou é executado iterativamente em cada fase para cada nível.

Variações dos valores dos parâmetros foram executadas na condução do experimento usando algoritmo Louvain. Os melhores resultados foram obtidos com os valores padrão de resolução (1,00000), número máximo de níveis em cada iteração (20) e número máximo de repetição em cada nível (50). O algoritmo Louvain escolhido executa o refinamento apenas no último nível.

Em seguida foram estabelecidos critérios para separação dos agrupamentos que não tratavam da temática de saúde, mesmo considerando uma visão ampla do domínio da área. Os critérios de exclusão constam de termos categorizados como pessoas, localizações geográficas específicas, além de organizações, eventos, espécies, tecnologias e campos de estudo não relacionados à saúde, como campos específicos da física ou matemática.

A Fase 1 resultou em uma rede de termos controlados-consumidor, característicos de um vocabulário de saúde, conectados por meio de relações semânticas como sinônimos, categorias e termos relacionado. Uma base de dados foi construída a partir dessa rede de forma a facilitar a organização de outras características relacionadas aos termos.

4.5 Fase 2 - Identificação de termos inéditos e validação humana (Objetivo 2)

O Objetivo 2 foi contemplado na Fase 2. Esta fase consta da análise de conteúdos não estruturados por meio de duas estratégias: recuperação de termos do UMLS e aplicação de técnicas de análise de textos para identificação de termos de saúde inéditos. A validação humana dos termos inéditos foi realizada para selecionar que termos candidatos poderiam compor o VSC (Figura 6).

As seções seguintes detalham o processo.

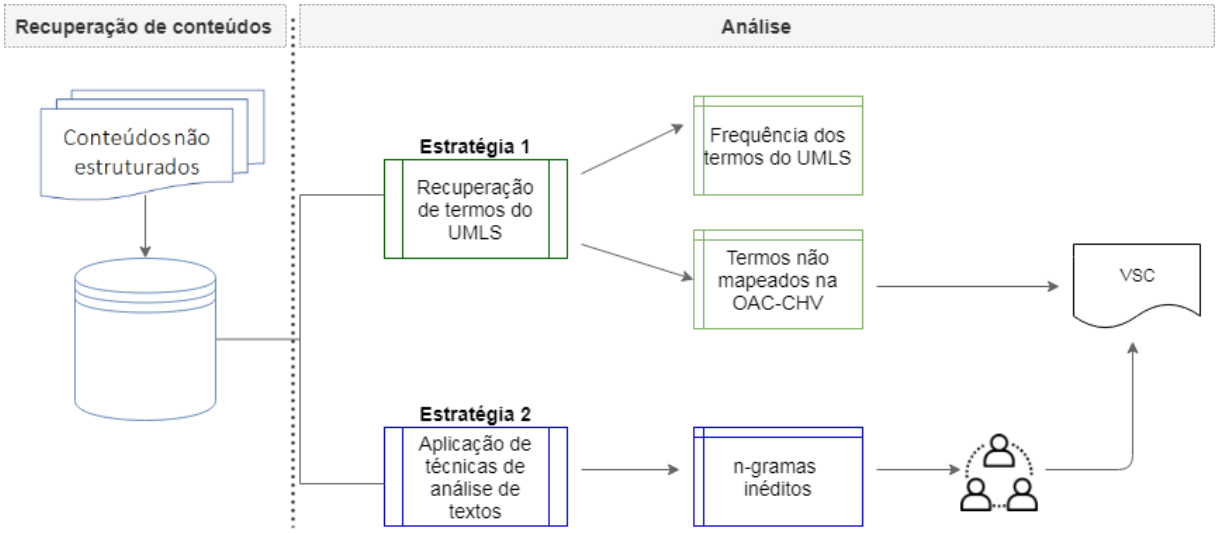


Figura 6. Processo para identificação de termos não mapeados e validação humana (Fase 2).

4.5.1 Coleta de conteúdos não estruturados

Os conteúdos não estruturados coletados são referentes a textos escritos para ou por consumidores em saúde.

Para compor esses conteúdos um script foi desenvolvido. Foram coletadas 209.458 notícias resumidas (NR) classificadas como ciência e saúde, emitidas por *feeds* Really Simple Syndication (RSS), publicadas entre 2013 e 2017. Adicionalmente, 1.924 notícias completas (NC) publicadas entre abril de 2012 e dezembro de 2017 pelo blog do Ministério da Saúde (blog.saude.gov.br) foram coletadas. Por meio deste blog também foi possível coletar 563 perguntas e respostas (PR) redigidas por associadas às notícias. Os exemplos de cada tipo estão indicados no Quadro 2.

Os conteúdos foram armazenados em uma aplicação para gerenciamento de banco de dados orientada a documento, o MongoDB versão 3.6 (mongodb.com). Esta aplicação foi escolhida por possibilitar realizar buscas em textos por termos exatos ou aproximados por meio da aplicação de algoritmos que executam a redução do termo ao seu radical (*stemming*) em idioma português.

Quadro 2. Exemplos de conteúdos referentes a notícias resumidas (NR), notícias completas (NC) e perguntas/respostas (PR).

Tipo	Exemplo
NR	“Aspirina diminui chances de desenvolvimento de câncer de cólon. Descoberta pode levar a outras formas de tratamento. Medicamento, no entanto, pode causar hemorragias."
NC	"O cancro mole é uma doença sexualmente transmissível causada por uma bactéria que causa feridas nos órgãos sexuais dos homens e das mulheres. Os primeiros sintomas como: dor de cabeça, febre e fraqueza aparecem de três a 13 dias após a relação sexual sem o uso da camisinha. Depois, surgem pequenas e dolorosas feridas com pus no pênis, na vagina ou no ânus, que aumentam progressivamente de tamanho e profundidade. O cancro mole ocorre com maior frequência nos homens e a sua existência aumenta a possibilidade de transmissão do vírus HIV. Para a diretora adjunta do Departamento de DST, Aids e Hepatites Virais do Ministério da Saúde, Adele Benzaken, a população deve ficar atenta a cuidados essenciais para não contrair a doença. A forma de evitar é pelo uso do preservativo em todas as relações sexuais do início ao fim da relação. Ou seja, não ter nenhum contato seja no sexo oral, anal e vaginal, sem o uso do preservativo. As IST’s, elas estão muito relacionadas, também, com o aumento no número de parceiros sexuais. Então, é importante o uso do preservativo e em alguns casos, redução no número de parceiros sexuais.” Se você percebeu ou sentiu qualquer sinal ou sintoma do cancro mole, é recomendado que procure um profissional de saúde mais próximo de casa, para o diagnóstico correto e a indicação do tratamento adequado. O cancro mole é tratado com o uso de antibióticos que são disponibilizados pelas unidades básicas de saúde de cada município."
PR	“Gostaria de saber se além da gestante, devem tomar a vacina dTPa outros adultos que terão contato com o bebê (tios, avós, babá) e, caso afirmativo, se para essas pessoas a vacina está disponível no sus. Obrigada"

4.5.2 Análise de conteúdos

As duas estratégias para análise dos conteúdos coletados constam de:

1. Recuperação de termos presentes nos vocabulários do UMLS nos conteúdos textuais, não obtidos na Fase 1;
2. Técnicas de análise de texto: aplicadas para recuperação de termos candidatos inéditos.

A Figura 7 mostra o fluxo para realização da estratégia 1.

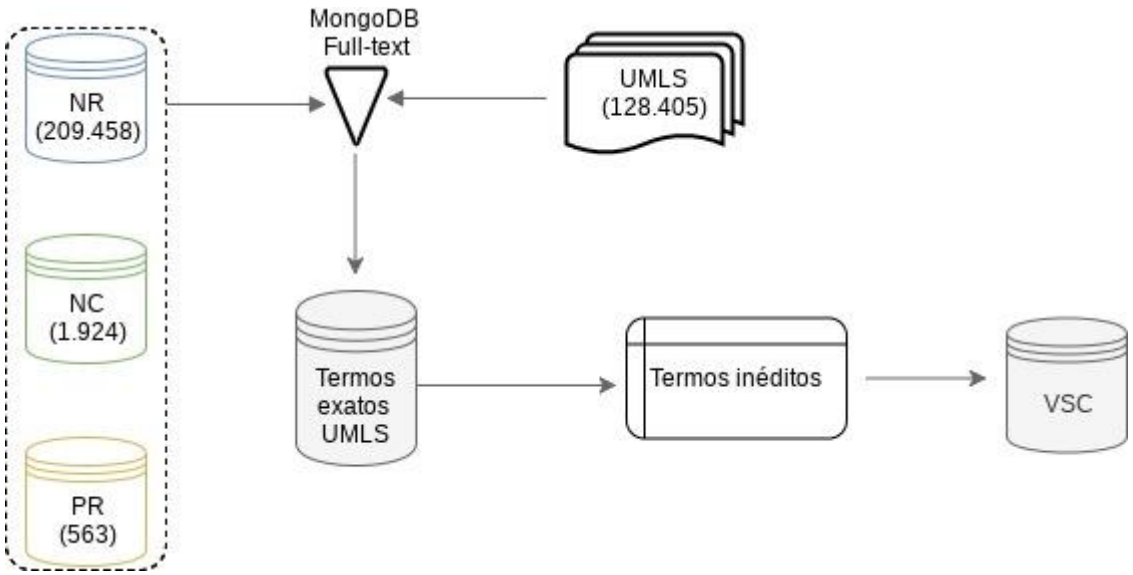


Figura 7. Recuperação de termos dos vocabulários UMLS nos conteúdos não estruturados (Estratégia 1).

Uma base de dados composta por termos distintos (termos preferidos e sinônimos) dos vocabulários do UMLS disponíveis em português foi gerada. A base foi utilizada para realizar uma busca pela ocorrência desses termos exatos nos conteúdos recuperados (NR, NC e PR). Essa tarefa foi realizada por meio da aplicação da técnica de busca *full-text*, suportada pelo software de gerenciamento de banco de dados MongoDB, que disponibiliza recursos para realizar buscas a partir de critérios estabelecidos pelo usuário porque suporta operações que realizam a busca em textos por meio de *text index* e do operador *$text.*[76]

A frequência de cada termo pertencente aos vocabulários do UMLS nos conteúdos não estruturados foi registrada. Essa ação possibilitou recuperar termos dos vocabulários do UMLS que ainda não pertencem à junção UMLS/OAC-CHV desenvolvida na Fase 1.

Quanto à estratégia 2, a Figura 8 mostra o esquema para a aplicação de técnicas de análise de textos, cuja finalidade é a identificação de termos inéditos.

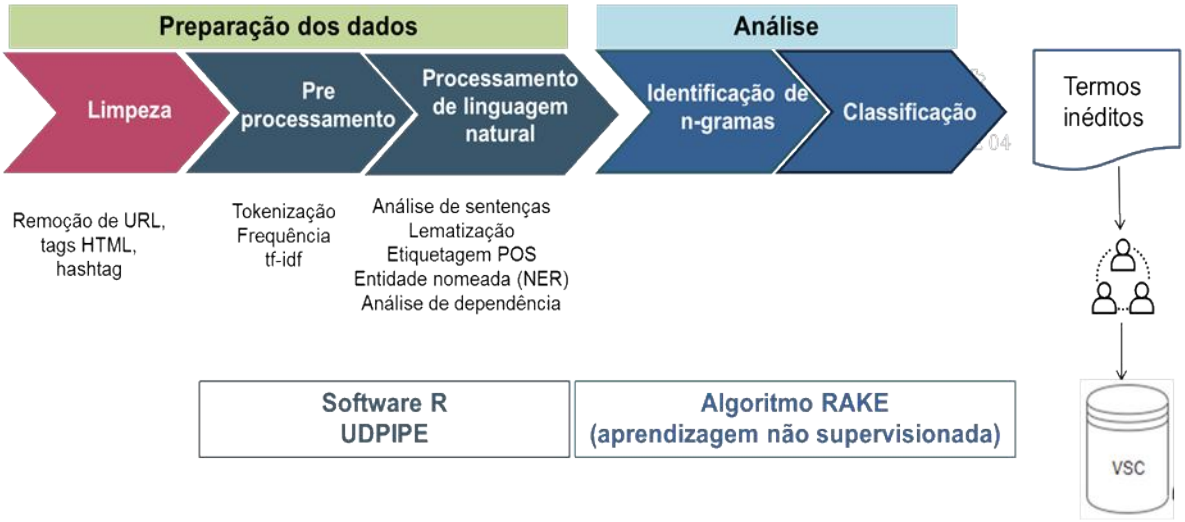


Figura 8. Técnicas de análise de textos aplicadas aos conteúdos não estruturados e validação humana (Estratégia 2).

A aplicação das técnicas de análise dos textos foi realizada por meio do software R (r-project.org). R é uma linguagem de programação e ambiente de software para análise estatística, representação gráfica e relatórios, que contém ferramentas importantes para a análise de textos.[77] A análise de textos envolve duas partes:

- Preparação de dados: são realizadas operações sobre palavras, como extração de tags HTML e URLs, e pré-processamento, referente à tokenização, *stemming* e remoção de *stopwords*, contagem de frequência e medidas como tf-idf. Outra técnica possível é a lematização, que utiliza um dicionário para substituir as palavras pela sua raiz morfológica. Ferramentas de processamento de linguagem natural também podem ser usadas para analisar sentenças, com a aplicação de técnicas de lematização, etiquetagem POS, reconhecimento de entidade nomeada e análise de dependência.
- Análise: aplicação de métodos de contagem considerando aspectos linguísticos, aprendizado de máquina supervisionado e não supervisionado.

Os conteúdos recuperados (209.459 NR, 1.924 NC e 533 PR) foram processados utilizando o pacote UDPipe[78] - modelo língua portuguesa.[79, 80]

UDPipe realiza automaticamente as etapas de extração de sentenças, tokenização, lematização e etiquetagem POS, o que possibilitou a identificação da classe gramatical de cada *token*.

Em seguida foram determinados critérios para a criação de um filtro baseado na etiquetagem POS. Para estabelecer um critério de seleção das classes gramaticais os termos dos vocabulários do UMLS foram analisados. A finalidade foi extrair termos com a estrutura linguística mais frequente dos termos do UMLS.

A análise foi realizada por meio da execução do algoritmo não supervisionado RAKE. Para isso foram realizados dois ensaios: um utilizando como filtro as quatro categorias gramaticais mais frequentes e outro utilizando as duas mais frequentes. RAKE retorna uma lista classificada de termos com n-gramas de cada conteúdo.

Uma lista única foi constituída com n-gramas classificados segundo o índice RAKE. A seguir foram removidos os termos exatos previamente mapeados nos vocabulários do UMLS.

De forma a estabelecer um critério de corte para os n-gramas identificados por RAKE uma análise foi realizada considerando fatores como o índice RAKE e a frequência do termo nos conteúdos não estruturados.

Para fins de obtenção de relações semânticas entre termos, como possíveis sinônimos, foram mapeados os termos candidatos que coocorrem com termos dos vocabulários do UMLS.

4.5.3 Validação humana

A aplicação do critério de corte resultou em uma lista com n-gramas a serem validados para compor o VSC.

O processo de validação foi composto pela avaliação dos n-gramas realizada por profissionais relacionados à saúde, seguida da análise de termos pela pesquisadora principal e análise estatística.

Uma interface web foi projetada e construída para facilitar e sistematizar a avaliação dos n-gramas, disponibilizada no endereço eletrônico (telemedicina.unifesp.br/projeto/vsc). Um chamado público à participação de avaliadores voluntários, especificamente, profissionais de saúde ou informática em

saúde, foi promovido. O chamado foi realizado por meio de listas de e-mail relacionadas à área de informática em saúde, redes sociais abertas e pelo ambiente colaborativo acessado por alunos do curso de especialização em informática em saúde da Universidade Aberta do Brasil/UNIFESP. O termo de consentimento livre e esclarecido foi disponibilizado para os participantes antes do processo de avaliação de termos (Apêndice 2 – Termo de Consentimento Livre e Esclarecido, p. 122). O período de avaliação ocorreu entre 22 de dezembro de 2018 e 10 de fevereiro de 2019. Não houve pagamento financeiro aos voluntários.

Inicialmente na avaliação uma página inicial composta por explicações sobre o projeto e instruções para a avaliação de termos foi apresentada aos voluntários. Em seguida os termos foram sorteados e apresentados aos avaliadores que deveriam indicar sua possível pertinência ao domínio da saúde de acordo com uma escala. A Figura 9 apresenta parte da interface de avaliação de termos.

Figura 9. Tela capturada de parte da interface web para avaliação de n-gramas disponibilizada aos avaliadores voluntários. Destaque para avaliação do termo "cirurgião vascular".

A quantidade de n-gramas a serem avaliados por cada avaliador não foi pré-determinada. Ao avaliar um n-grama coube ao avaliador a opção de prosseguir para o próximo item. Garantiu-se, entretanto, que cada n-grama fosse avaliado ao menos uma vez.

Para análise final a escala foi simplificada e considerou-se validados os n-gramas classificados como "é saúde" ou "pode ser (saúde)" e não validados os classificados como "não é saúde" ou "não faz sentido" (n-gramas inválidos) (Figura

9).

Na continuidade do processo, cada termo foi avaliado de acordo com as regras:

1. Todos os termos foram reavaliados pela pesquisadora principal por meio da escala simplificada considerando os valores “validado” ou “não validado”.
2. Em caso de dúvida da classificação final foram consultados os vocabulários do UMLS, que subsidiaram a validação de variantes dos termos, a exemplo de “Informes Anuais” [DeCS] e “Informe” [VSC]. Outro item auxiliar foi buscar o significado do termo em vocabulários disponíveis na web.

   3.Os casos de divergência entre as avaliações foram dirimidos por meio da consulta a uma profissional de saúde voluntária, graduada em farmácia em 2008 e especialista em farmácia hospitalar há 10 anos. Esta especialista utiliza rotineiramente listas de termos de saúde em seu exercício profissional.

Desta forma foi possível obter o padrão-ouro de termos inéditos para integrar o VSC que foi utilizado para obtenção das métricas de avaliação de desempenho do algoritmo RAKE.

Medidas de concordância entre as duas avaliações foram realizadas, como o percentual de concordância e coeficientes de concordância, como o Kappa de Cohen (κ), uma medida estatística usada para medir a variação entre as avaliações dos observadores.[81] Esses cálculos foram realizados por meio do pacote REL [82] disponível para o software R.

4.5.4 Avaliação de desempenho do algoritmo RAKE

Dado que um algoritmo baseado no método não supervisionado foi utilizado, o desempenho do algoritmo RAKE foi avaliado de acordo com o padrão-ouro estabelecido por meio da validação humana.

Medidas estatísticas foram realizadas para fins de verificação da capacidade

do algoritmo RAKE de recuperar termos pertencentes ao domínio da saúde válidos. O algoritmo RAKE foi projetado para recuperar palavras-chave em textos diversos, não especificamente para uma determinada área, o que dificulta a tarefa de encontrar termos válidos específicos da saúde.

O desempenho de classificadores automáticos é analisado por meio das curvas *receiver operating characteristic* (ROC) e precisão-revocação (PR), que são ferramentas de diagnóstico que auxiliam na interpretação da previsão probabilística para problemas de modelagem preditiva de classificação binária.[83]

A curva ROC exprime o relacionamento entre a taxa de sensibilidade e 1-especificidade para um modelo preditivo usando diferentes limites de probabilidade. De forma análoga a curva PR resume o relacionamento entre a taxa positiva verdadeira e o valor preditivo positivo do modelo.[84]

Métricas padrão utilizadas na análise de classificadores automáticos como sensibilidade, especificidade e área sobre a curva ROC (AUC-ROC) foram realizadas. A AUC-ROC é uma medida importante para fins de comparação com outros estudos.

Além disso, medidas utilizadas em recuperação da informação como precisão, revocação e *F-measure* foram realizadas nesse estudo. De fato, estas medidas se mostraram mais adequadas para avaliar o conjunto de termos validados.[83] *F-measure* é definida como a média harmônica entre precisão e revocação.

4.6 Aplicação computacional (Objetivo 3)

Esta seção apresenta o desenvolvimento de uma aplicação computacional para acesso ao VSC suportada pela representação semântica dos dados. O conteúdo e estrutura do VSC, previamente mapeados por meio de uma rede complexa, foram padronizados de acordo com ontologias específicas selecionadas a partir de uma pesquisa prévia descrita nas próximas seções. Um modelo de dados VSC-RDF foi construído e armazenado em um servidor de banco de dados.

4.6.1 Seleção de ontologias

Para a criação do VSC em RDF ontologias específicas devem suportar a formalização do vocabulário por meio de classes e propriedades para representação dos dados, assim como de seus metadados.

De acordo com a web semântica o modelo RDF estende a relação entre os identificadores URI por meio do mapeamento de seus relacionamentos. Para tal finalidade três componentes básicos foram definidos: sujeito, predicado e objeto, que formam as triplas. O sujeito é um recurso que pode conter a URI, o predicado descreve as propriedades da classe e o objeto consiste no valor da propriedade. As ontologias estabelecem a padronização dos predicados, de forma a serem reutilizados em outros modelos de dados.[85]

As ontologias foram selecionadas a partir da pesquisa realizada em bases bibliográficas e sites responsáveis pela publicação destas (w3.org/TR/?status=cr e bioportal.bioontology.org/ontologies) de forma a verificar a ocorrência de modelos que suportassem a representação de dados e metadados.

Ontologias orientadas para a representação de vocabulários em geral e específicas para CHVs estão disponíveis e serão apresentadas na sequência.

O Simple Knowledge Organization System (SKOS) [86] é a ontologia base para representação de esquemas de organização de conhecimento. Trata-se de uma iniciativa do W3C que possibilita a representação do conhecimento em um formato legível por máquina baseado em grafos RDF.[20] SKOS fornece um modelo para representar a estrutura básica e o conteúdo de esquemas de conhecimento como dicionários, taxonomias, listas de classificação e outros esquemas.[87]

SKOS propõe que conceitos sejam formalizados como uma classe e representado formalmente como uma *skos:Concept*. Os conceitos podem ser identificados com URIs, rotulados com cadeias lexicais em uma ou mais linguagens naturais, documentadas com vários tipos de notas, semanticamente relacionadas entre si em hierarquias informais e redes de associação e agregadas em esquemas conceituais. A relação entre o termo preferido (representado por *skos: prefLabel*) e um termo leigo é formalizado por meio do uso de etiquetas alternativas SKOS (*skos: altLabel* ou *skos: hiddenLabel*).[88]

Uma ontologia específica para representação do CHV foi encontrada [24], porém, segundo a própria descrição trata-se de uma implementação baseada na ontologia SKOS. Outro aspecto importante é que as propriedades associadas a essa ontologia não suportam por completo os elementos do modelo de dados e relações deste estudo.

SKOS, porém, não é suficiente para representar os metadados de proveniência necessários para identificar os termos coletados ou extraídos de vocabulários controlados e os de domínio do consumidor, o que pode ser um item importante para o uso dos CHVs. Para suportar os metadados a ontologia MuEVo [89] foi proposta, baseada em SKOS e PROV Ontology, uma recomendação do W3C para formalizar informações de proveniência. O CHV em idioma francês[22] foi formalizado por meio da ontologia MuEVo.

Todavia, a ontologia MuEVo não contempla, de forma direta, as descrições essenciais para rastrear a proveniência, a criação e a versão dos recursos da web. Para esse fim a ontologia Provenance, Authoring and Versioning ontology (PAV) (purl.org/pav) foi publicada. PAV distingue os agentes responsáveis pelo conteúdo, além da proveniência dos recursos originários que foram importados, transformados e consumidos.[90]

Após análise das ontologias concluiu-se que as ontologias SKOS e PAV são suficientes para representar os dados do VSC. SKOS possibilita a representação de conceitos, termos preferidos, sinônimos, termos relacionados, categorias e definição. PAV fornece a representação para a origem do termo, data de atualização ou acesso, versão do recurso e autoria.

4.6.2 VSC-RDF

O conteúdo e estrutura dos termos resultantes das Fases 1 e 2 compuseram uma base de dados utilizada para compor o VSC em RDF, o modelo padrão para intercâmbio de dados. Esse modelo possibilita a troca de dados, busca de informações de saúde e integração outros projetos.

A estrutura consolidada é referente ao mapeamento dos relacionamentos semânticos (sinônimos, hiperônimos e termos relacionados), identificadores (ID, CUI,

DUI), definições extraídas do DeCS e origem do termo.

Recursos disponibilizados pelo pacote rdflib [91] disponível para o software R contribuíram para a construção da representação do VSC de acordo com o padrão RDF. Este pacote fornece uma interface para executar tarefas comuns em modelos de dados RDF, como leitura e escrita, além de converter os dados entre as várias serializações, incluindo rdfxml, nquads, ntriples e json-ld.[92, 93] Outro fator importante é a execução de consultas por meio da linguagem SPARQL que possibilita recuperar os dados do grafo RDF.[94]

Posteriormente as triplas foram inseridas em um arquivo local e em um banco de dados criado para este projeto.

4.6.3 Interface de acesso ao VSC

Uma interface web foi projetada para suportar a estrutura do VSC e possibilitar o acesso ao conteúdo construído. As tecnologias para desenvolvimento web usadas constam de NodeJS (nodejs.org) em conjunto com o framework para aplicações web Express (expressjs.com) e com o sistema gerenciador de banco de dados MySQL (mysql.com).

Ao realizar uma busca por termo ao usuário é apresentado o conteúdo referente a sinônimos, identificadores, categorias, termos relacionados e definição. A navegação pode ser realizada por meio dos termos relacionados. Outra possibilidade é a exploração do conteúdo diretamente pela rede complexa.

5 RESULTADOS

As próximas seções expõem os resultados desta tese de acordo com os objetivos específicos previamente definidos.

5.1 Revisão da literatura

A Tabela 1 mostra o resultado da estratégia de busca, aplicação dos critérios de inclusão e exclusão e seleção sequencial de leitura. A estratégia de busca resultou em 1.070 no Google Scholar, porém, devido à grande quantidade foram analisados apenas os 200 primeiros artigos, que são os mais relevantes segundo o critério de classificação do Google. Dentre os artigos da base Google Scholar selecionados para leitura completa, um não estava em idioma inglês e o outro era apenas de uma duplicação de um artigo selecionado. Da base PubMed um dos artigos foi excluído na fase de leitura de títulos por ser referente a este estudo.

Tabela 1. Quantidade de artigos selecionados por meio da estratégia de busca, aplicação de critérios e seleção por leitura.

	Critérios			Seleção			
Base	**Inclusão**	**Exclusão**	**Duplicações**	**Títulos**	**Resumo**	**Texto**	**Análise**
PubMed	173	73	0	19	16	15	13
Periódicos Capes	181	71	8	62	25	4	3
Google Scholar	200	33	12	21	3	1	1
Total				102	44	20	15

Os 15 artigos selecionados para leitura foram analisados. Devido à variabilidade apresentadas nos estudos em relação aos objetivos e métodos, um resumo de cada artigo foi esquematizado, enfocando aspectos como: que bases de dados publicadas foram utilizadas, uma breve descrição, que métricas foram obtidas e os principais achados. Essa análise foi organizada e apresentada no Quadro 4 disponível em anexo (Anexo 1 – Síntese dos estudos recuperados na revisão da literatura, p. 125).

A análise resultou que 80% (12/15) dos estudos foram desenvolvidos entre

2016 e 2018, o que indicou um novo ciclo de importância com possibilidades de desenvolvimento de métodos e aplicações dos vocabulários de saúde do consumidor, com a condução de pesquisas ativas que objetivam obter e ligar termos do consumidor aos técnicos, assim como torná-los disponíveis para utilização do consumidor ao acessar registros médicos ou para o desenvolvimento de materiais educacionais.[95]

Apesar do conteúdo desses vocabulários estarem voltados para o consumidor, apenas 47% (7/15) dos estudos utilizam conteúdos escritos por ou para consumidores.[23, 34, 9699] De fato, por se tratarem de conteúdos com menor grau de padronização que os registros eletrônicos de saúde, as tarefas de análise e extração automática de termos válidos é dificultada, porém devido à disponibilização de ferramentas que auxiliam a análise de textos, os conteúdos gerados por consumidores tornaram-se o foco dos estudos mais recentes.

Outro aspecto importante é que o OAC-CHV tornou-se uma base de conhecimento básica para novos experimentos, utilizada em 60% (9/15) dos estudos. Foram identificadas variadas utilizações da OAC-CHV. Para fins de síntese das contribuições dos estudos para a evolução do desenvolvimento dos vocabulários foi elaborada a Figura 10.

Métodos e ferramentas para extração automática de termos foram aplicados a conteúdos recuperados de redes sociais, como Tumblr [96], Yahoo Answers [34, 96], Twitter [99], além de fóruns [22, 9799] e registros eletrônicos do paciente (*electronic health record* - EHR).[99, 101] As técnicas de RAT, tf-idf e c-value foram aplicadas em 5/8 artigos que as utilizaram.[22, 33, 34, 99, 101] A análise linguística, por meio de padrões sintáticos, como etiquetagem POS também foi aplicada em 3/8.[22, 33, 34] Outras técnicas aplicadas foram modelo de tópicos [33, 98, 101] e *word embedding* (word2vec e outros análogos).[99, 99, 101] Os estudos em geral aplicaram várias técnicas simultaneamente.

A classificação de termos foi realizada por meio de aprendizagem de máquina. Sistemas baseados em processamento de linguagem natural e aprendizagem não supervisionada foram descritos e mostraram-se eficientes na tarefa de identificação de termos inéditos.[33, 99, 101] Um método de aprendizagem supervisionada para mapeamento de sinônimos também foi descrito [95], assim como

uma técnica que refina os termos extraídos por *word embedding* por meio de uma abordagem supervisionada.[99]

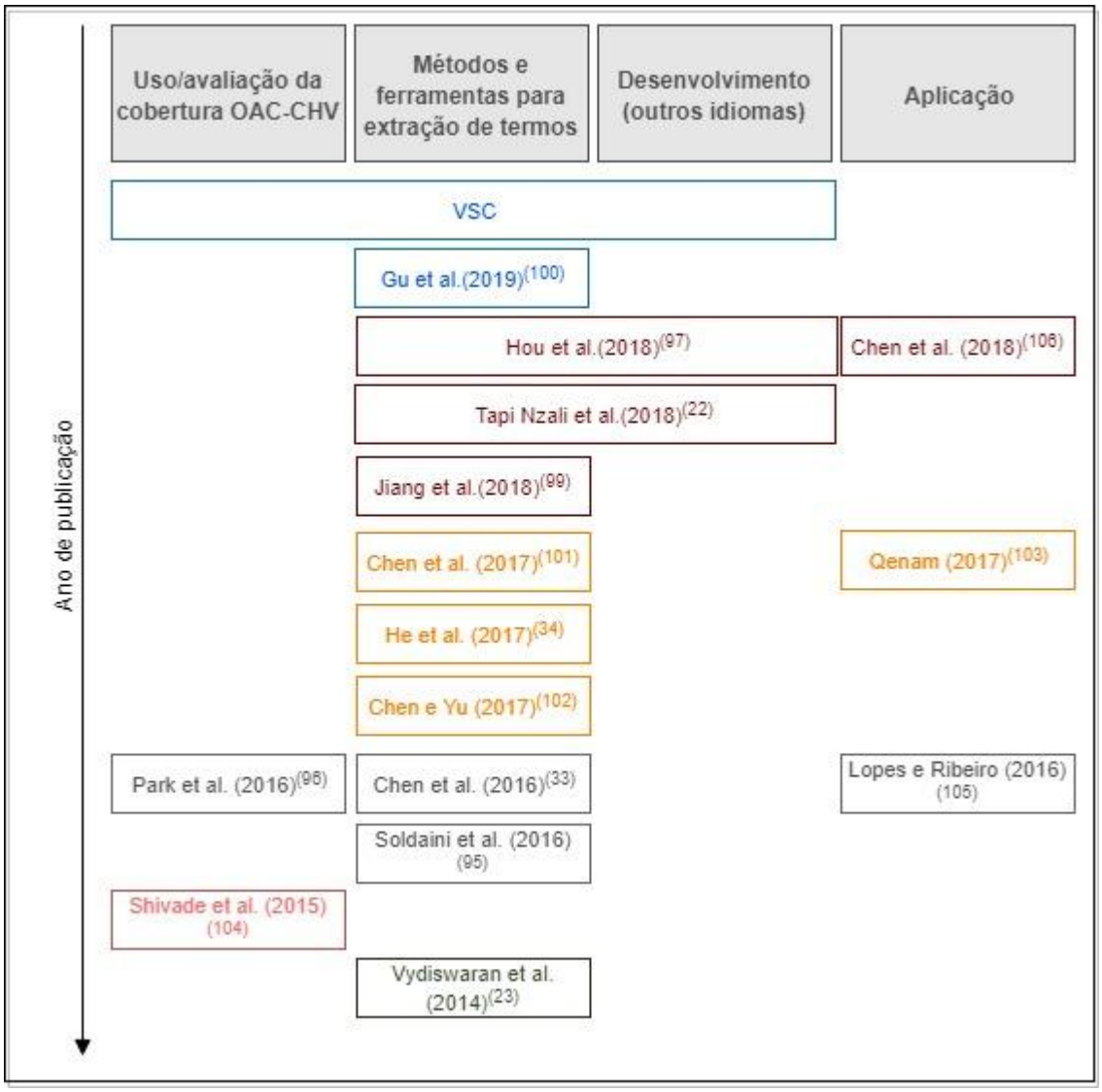


Figura 10. Síntese da colaboração dos estudos baseados na OAC-CHV com a inclusão do presente estudo.

O esforço humano representa a base para avaliação dos métodos automáticos [33, 101] e para extração manual de termos.[97] A necessidade de validação humana para comparação de resultados, considerada o padrão-ouro, foi contemplada por estudos que desenvolveram métodos automáticos supervisionados ou não-supervisionados para recuperação e classificação de termos utilizando métricas como precisão*, F-measure* e área sob a curva ROC (AUC-ROC).[33, 99, 101]

Alguns estudos baseiam-se na verificação da cobertura da OAC-CHV diante de conteúdos gerados por ou para consumidores em saúde.[34, 103, 104] Estes estudos são a base inicial para o desenvolvimento de aplicações. Apenas uma aplicação efetiva de vocabulários voltada para a recuperação de informação na web foi encontrada.[105] Outra aplicação foi encontrada, porém em um estudo não recuperado pelo método aplicado. Chen et al.[106] descreveram o desenvolvimento de uma aplicação baseada no reconhecimento de termos técnicos pela qual se ofereciam sinônimos e definição de termos encontrados em notas de prontuário eletrônico do paciente, por meio de uma interface. Este estudo foi a continuação de uma estrutura prévia criada pelos autores.

O estudo que descreveu o desenvolvimento de CHV para o idioma francês [22] utilizou a ontologia SKOS para representação do conhecimento. Observa-se que, apesar de única, trata-se de uma iniciativa importante para que os vocabulários possam utilizar dos conceitos da web semântica.

A evolução do desenvolvimento dos modelos de vocabulário tende à utilização de métodos automatizados que recuperem e elejam termos candidatos que poderão compor a base final dos vocabulários. Sousa[17] mostrou que três entre seis artigos publicados sobre o desenvolvimento de modelos de vocabulários utilizaram método manual para a recuperação de termos. Nos últimos cinco anos apenas um estudo recente e incipiente aplicou esse método para fins de recuperação de conteúdo[27], o que indicia uma evolução na aplicação de técnicas atuais para o desenvolvimento de vocabulários.

5.2 Construção da rede complexa de termos (Objetivo 1)

Iniciamos esta seção apresentando o resultado da coleta e extração de dados de fontes estruturadas e, na sequência, a estruturação e caracterização da rede complexa composta por termos gerados por vocabulários controlados e do consumidor, que cumprem o Objetivo 1.

5.2.1 Coleta e extração de termos de fontes de dados estruturados

O esquema de coleta e extração de termos dos vocabulários controlados está representado na Figura 11. A intersecção das bases OAC-CHV (158.519 termos) e UMLS (128.405 termos) por meio do CUI resultou em 49.809 termos de domínio do consumidor em saúde. Desses, 24.131 são referentes termos preferidos (denominados descritores no MeSH/DeCS), sendo 16.148 pertencentes à base MeSH. Cerca de 50% (32.701/65.150) dos termos do MeSH foram coletados.

Os termos (49.809) foram pesquisados no serviço DeCS para extrair as relações semânticas. Foram recuperados 78% dos termos preferidos pesquisados (12.613/16.148). Para cada conceito foram recuperados os sinônimos (40.468), categorias (20) e outros termos preferidos relacionados por hierarquia (árvore MeSH) e por coocorrência (6.869), com sobreposição de termos. A pesquisa por sinônimos ocorreu pelo fato de que o DeCS disponibiliza mais sinônimos para alguns termos que os dicionários do UMLS. Uma base de dados foi composta por 66.306 termos controlados, identificadores e suas relações semânticas.

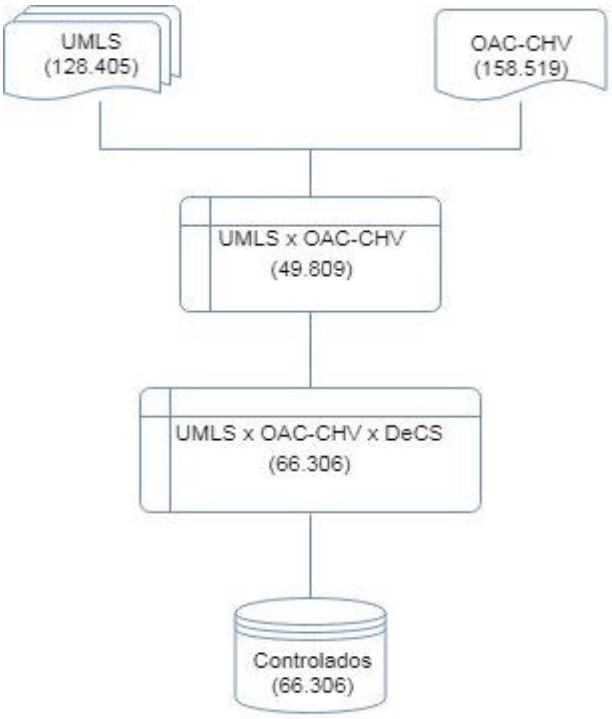


Figura 11. Coleta e extração de dados controlados das fontes estruturadas.

O serviço DBpedia SpotLight foi utilizado para anotar os termos controlados (66.306) e respectivas URIs. Foram anotados 40.851 termos dos vocabulários controlados. Destes, 6.048 eram referentes a termos exatos ou flexionados por número (singular/plural) considerando termos preferidos e sinônimos (9,1%). Os termos restantes (34.803) foram resultantes da anotação semântica, que é executada pelo DBpedia SpotLight ao não encontrar o termo exato ou flexionado. A anotação semântica é importante para a extração de sinônimos ou partes dos termos de busca.

Para cada URI foram extraídos a categoria (13), sinônimos (20.720), termos relacionados por coocorrência (110.465) e assuntos (33.636). Há sobreposição entre os termos. Desta forma a base do consumidor foi composta por 179.198 termos únicos e suas relações. A Figura 12 mostra a síntese do processo de extração de dados realizada.

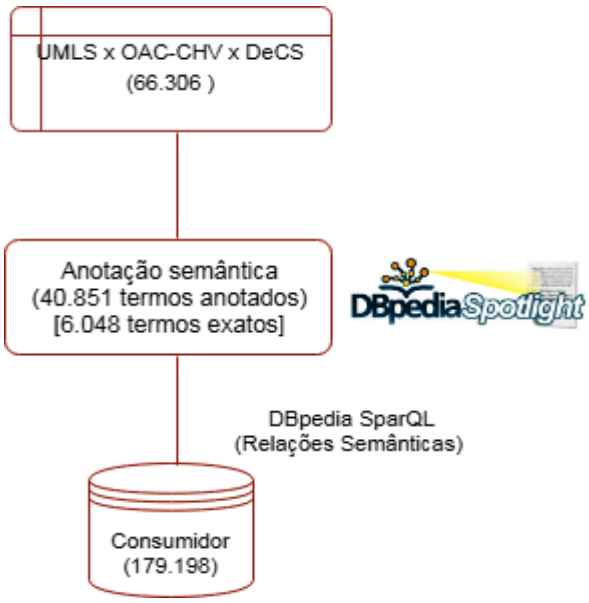


Figura 12. Processo de extração de dados para construção da base do consumidor.

Na sequência arquivos compostos por termos de ambas as bases foram organizados para viabilizar a construção de redes complexas.

5.2.2 Construção e caracterização das redes complexas

As redes de termos controlados e do consumidor foram construídas separadamente. Para fins de comparação o algoritmo Louvain foi executado, o que resultou na obtenção de agrupamentos cuja estrutura possibilita verificar a distribuição de tópicos mapeados por meio dos termos representativos de cada agrupamento.

A Tabela 2 apresenta os 10 maiores agrupamentos de cada rede complexa. Os termos são representativos do agrupamento. Verifica-se a predominância na rede de termos controlados dos termos categorizados como “compostos químicos e drogas” (18.2%) e “anatomia” (8.9%). Em relação à rede de termos do consumidor observa-se a predominância dos termos categorizados como “organismos” (25,4%) e “localização geográfica” (6,5%).

Tabela 2. Distribuição de frequências (F) dos clusters.

Controlados (grau médio: ~4,6)		Consumidor (grau médio: ~5,6)	
Descrição	**F (%)**	**Descrição**	**F (%)**
estruturas animais	8,9	lugar	16,5
aminoácidos, peptídeos e proteínas[1]	6,7	doenças	15,9
assistência à saúde	6,4	plantas[2]	10,7
meio ambiente e saúde pública	6,2	cultura/costumes	9,3
compostos organicos[1]	4,5	substância/processos químicos e físicos	9,2
enzimas e coenzimas[1]	3,9	organismos/animais[2]	5,0
doenças	3,9	organismos/anfíbios e pássaros[2]	3,9
comportamento humano	3,3	organismos/mamíferos[2]	3,5
ações químicas e utilizações[1]	3,1	regiões administrativas[2]	3,4
viroses	3,0	organismos/insetos	2,3
Total	**49,9**	**Total**	**79,7**

[1]Clusters que agregam termos caracterizados pela categoria compostos químicos e drogas.
[2]Clusters que agregam termos caracterizados pela categoria organismos.

O software Pajek possibilitou unir as redes por meio dos vértices e unificar os termos presentes em ambas, o que resultou em uma rede com 228.124 vértices. Foram mapeados 10.413 termos comuns, ou 15,7% dos termos controlados. Considerando apenas os termos preferidos (24.131) foram contabilizados 8.817 (36,5%).

A densidade da rede de termos controlados-consumidor foi 2,4 x 10^{-5} e o número de ligações entre os vértices (grau) médio foi de 6,3 com amplitude [1 - 11.905]. Cerca de 39% dos termos têm grau 1 e 90,8% tem grau máximo 10 e são

termos em sua maioria representados por compostos químicos e organismos.

Para fins de caracterização da rede de termos controlados-consumidor foram calculadas medidas de centralidade, apresentadas na Tabela 3. As medidas são importantes para mensurar os aspectos referentes à importância do vértice na rede.

Os resultados da Tabela 3 estão de acordo com o esperado para uma rede baseada em dados característicos da área de saúde considerando a classificação do MeSH/DeCS. As medidas realizadas referem-se aos clusters identificados e caracterizam o conteúdo da rede.

Tabela 3. Medidas de centralidade da rede de termos controlados-consumidor.

Grau [1 – 11.905] 5,94 ± 42,00		Proximity Prestige Input [0,000 – 0,134] 0,057 ± 0,041		Betweeness [0,0000 – 0,0250] 0,0000±0.0,0002	
angiosperma	11,905	organismos	0,134	#animalia	0,0250
localizações geográficas	3,415	doenças	0,130	#brasil	0,0203
#mollusca	3,173	#animalia	0,128	#saúde ambiental	0,0188
Mamíferos	3,055	#células	0,128	#plantae	0,0169
pássaros	2,399	#genética	0,128	#estados unidos	0,0151
anfíbios	1,747	#brasil	0,128	#compostos químicos	0,0139
peixes	1,672	#biologia	0,128	compostos químicos e drogas	0,0138
#brasil	1,669	#saúde	0,127	#química	0,0130
#compostos químicos	1,597	#água	0,127	#áfrica	0,0119
#estados unidos	1,533	#bactéria	0,126	#doenças	0,0108

Obs.: Termos iniciados por # pertencem às redes de termos controlados e do consumidor.

O termo "angiosperma" é categorizado como "organismos", assim como "moluscos", "mamíferos", "pássaros", "anfíbios" e "peixes". O alto valor associado ao grau de entrada/saída mostra que esses são termos que os mais interagem com outros termos e que, num modelo de rede, os vértices com maior grau podem ser associados a categorias.

Proximity prestige indica a importância ou influência dos vértices em relação à rede.[47, 48] De fato, essa medida indica que na rede controlados-consumidor os agrupamentos de termos categorizados como organismos, doenças e componentes químicos são os que ocupam a posição mais centralizada na rede.

Vértices com menor grau ou menor valor de *proximity prestige* são referentes

a ligações mais raras ou mais periféricas, que podem ser acessíveis somente a partir de um nível diferenciado de navegação. Estes vértices, porém, são importantes para a exploração da rede porque são referentes, por exemplo, a termos de saúde mais raros.

Betweenness nos mostra os vértices que fazem a comunicação com outros grupos de vértices.[47, 48] Observa-se que os vértices com maior valor são pertencentes à ambas as redes, o que é indicado por "#" na Tabela 3. Estes são os termos que unem os agrupamentos, por exemplo, de doenças a componentes químicos.

Para visualização do relacionamento entre os termos componentes da rede controlados-consumidor foi induzida uma sub-rede de 800 vértices (Figura 13). Observa-se que o modelo em rede proposto possibilitou explorar o relacionamento entre termos de categorias diferentes.

A Figura 14 detalha os vértices associados a “células”, evidenciando a relação entre as categorias.

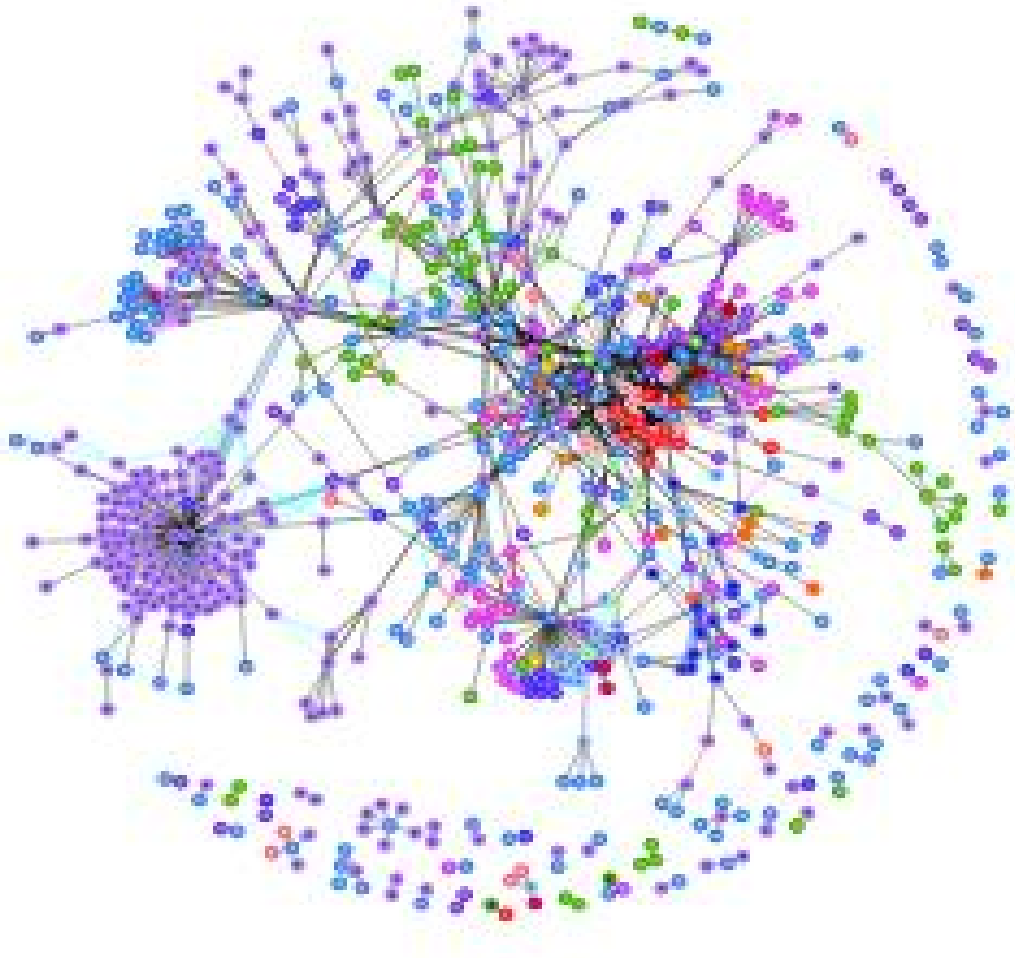


Figura 13. Sub-rede com 800 vértices evidenciando a categoria dos vértices. Doenças (verde), anatomia (azul), compostos químicos e drogas (roxo), localizações geográficas (vermelho), técnicas e equipamentos (rosa). Versão interativa disponível em https://bit.ly/2qHMiFY.

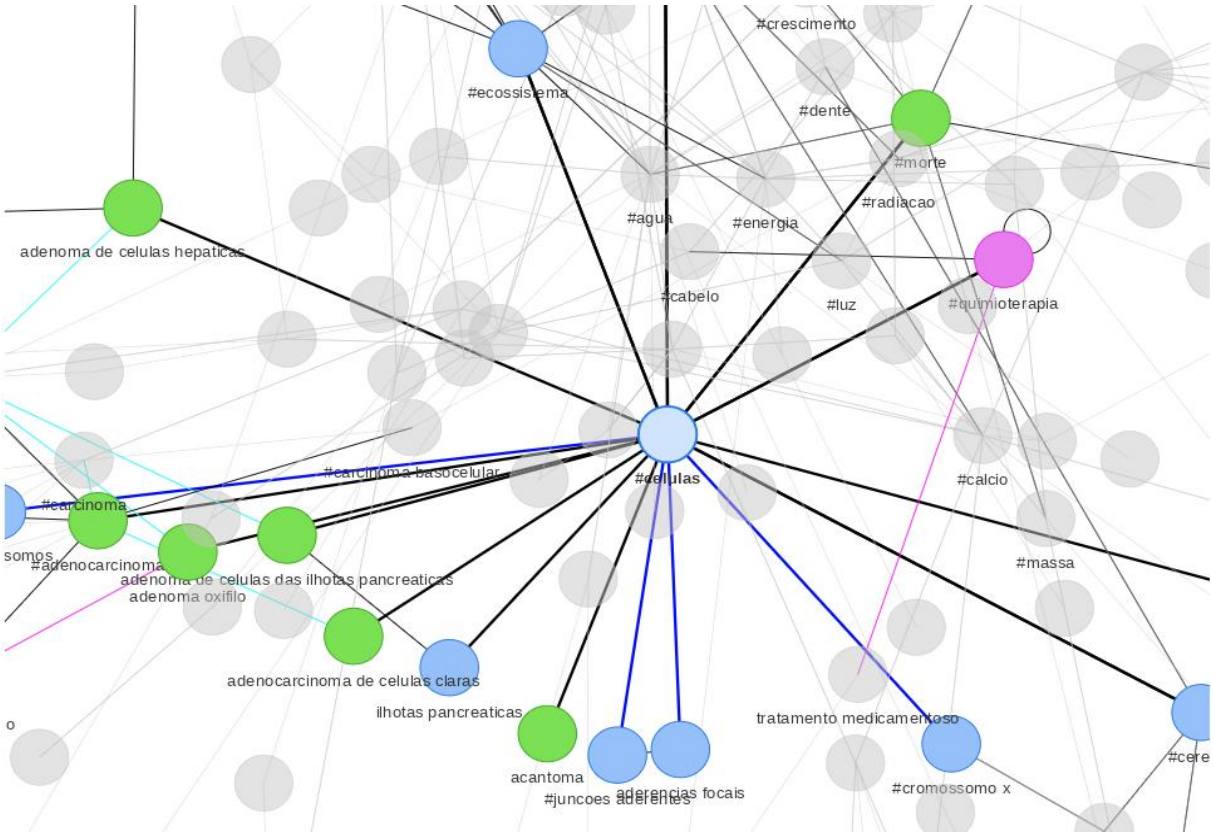


Figura 14. Vértices relacionados à “célula”, evidenciando as categorias. Versão em alta resolução disponível em https://bit.ly/2OE01ad.

O modelo de rede proposto foi suportado pelos relacionamentos semânticos, como sinônimos, hiperônimos (categorias) e termos relacionados. A Figura 15 evidencia essa estrutura por meio da sub-rede induzida pelo termo preferido “influenza”. As Figuras 16 e 17 evidenciam detalhes da sub-rede da Figura 15, indicados pelos rótulos dos vértices.

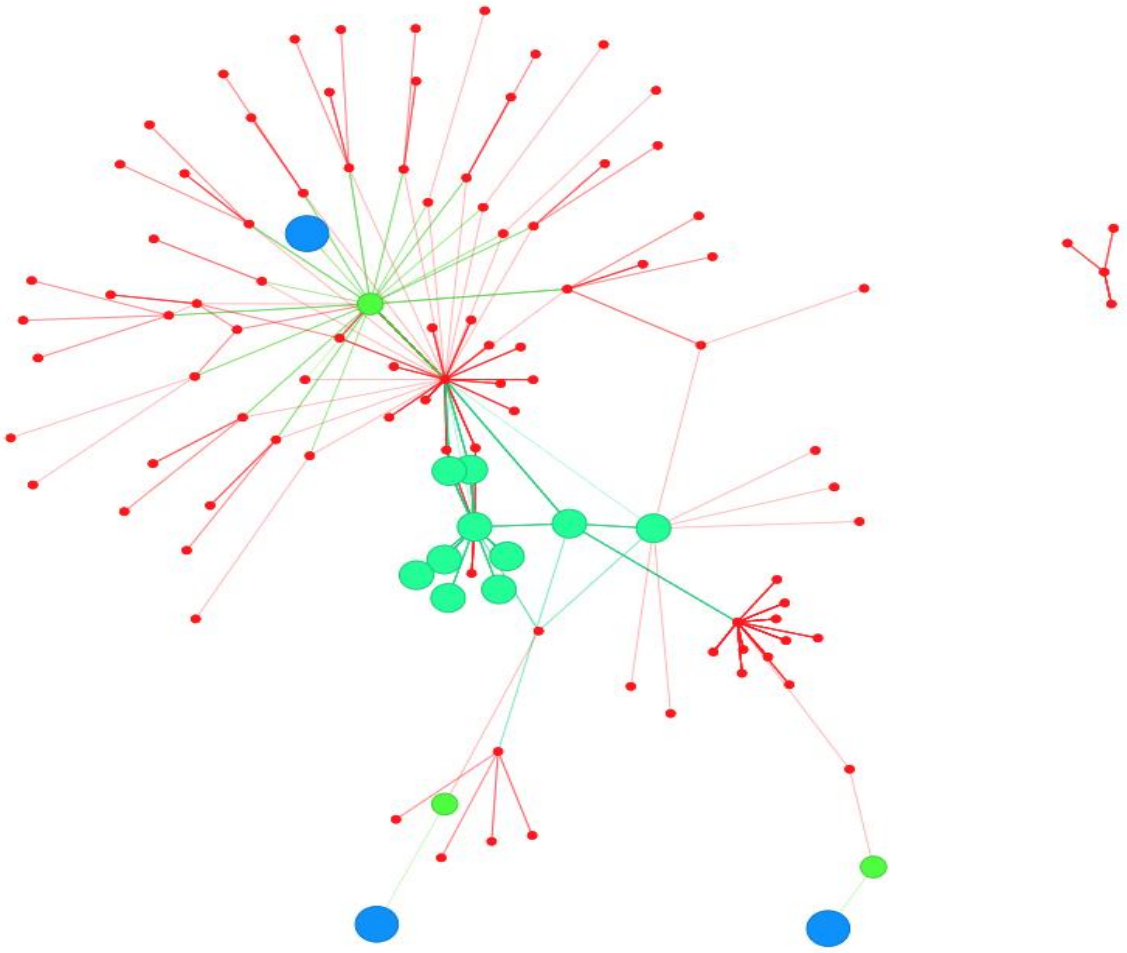


Figura 15. Exemplo de sub-rede induzida pelo descritor "influenza". Os vértices (azul) representam as categorias dos termos ligados a eles. Os vértices e arestas em verde (centro) representam os sinônimos e os vértices e arestas em vermelho representam os termos relacionados. Versão interativa em https://bit.ly/2zFjnGy.

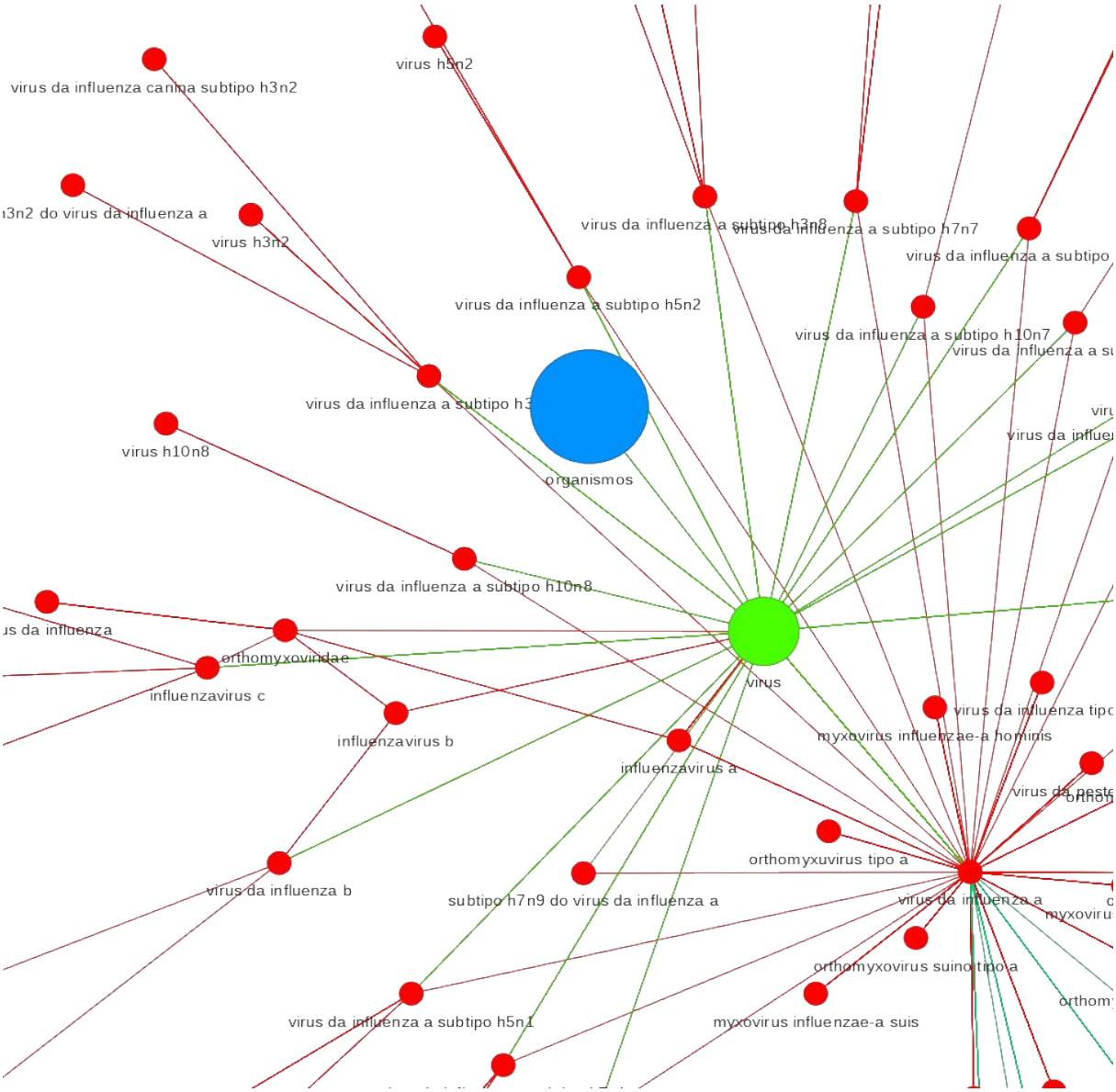


Figura 16. Detalhe da sub-rede induzida pelo termo preferido "influenza". "Organismos" (azul) é a categoria de "vírus" (verde) e "influenza vírus a" é um termo relacionado a "vírus" (vermelho). Versão em alta resolução disponível em https://bit.ly/2Dh84aL.

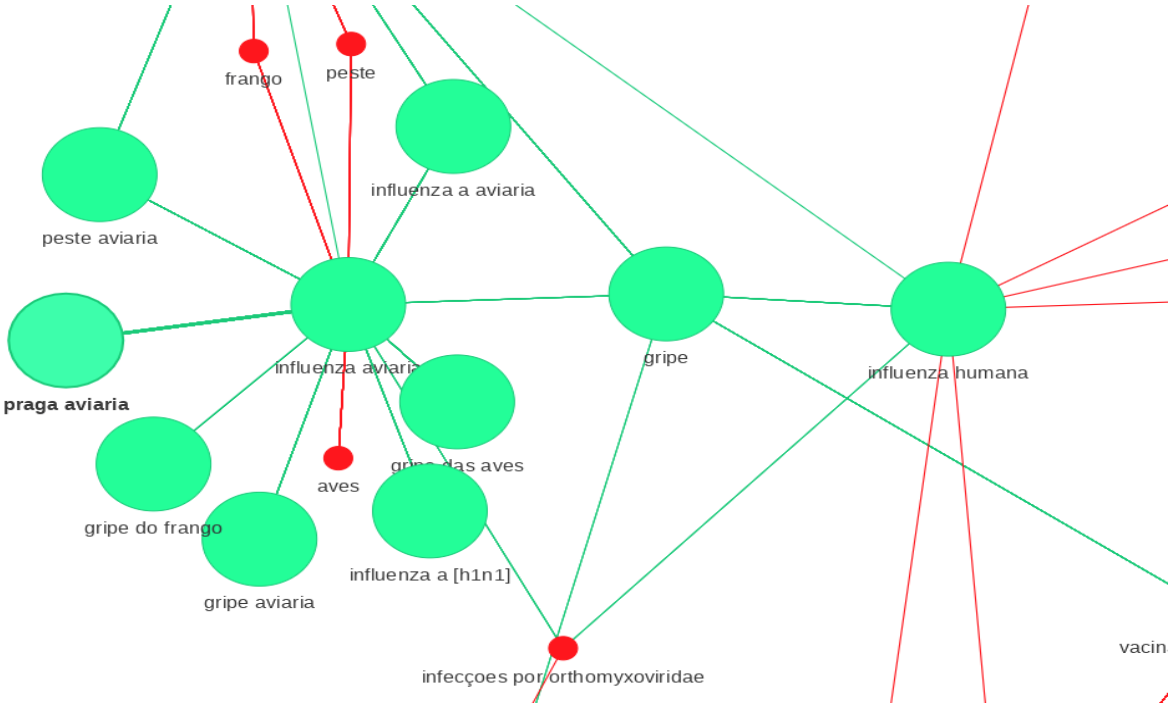


Figura 17. Detalhe da sub-rede induzida pelo termo preferido “influenza”, que mostra os sinônimos associados aos termos “gripe” e “influenza humana”.

5.2.3 Separação de termos

A execução do algoritmo Louvain detectou automaticamente 300 comunidades. A modularidade da rede foi calculada em 0,771805. Redes com alta modularidade apresentam uma densa conexão entre os vértices da mesma comunidade e conexões esparsas entre os outros vértices.

Após a aplicação dos critérios de exclusão restaram 165 comunidades. Devido à dificuldade de avaliar uma comunidade referente a "compostos químicos", pela heterogeneidade dos termos, o algoritmo Louvain foi executado novamente. O critério de exclusão foi aplicado para cada comunidade. A exclusão foi realizada somente para os termos recuperados da DBpedia.

Para efeito de visualização foram selecionados vértices com grau maior que 50. A Figura 18 apresenta a estrutura dos vértices das dez maiores comunidades identificadas por meio do termo representativo.

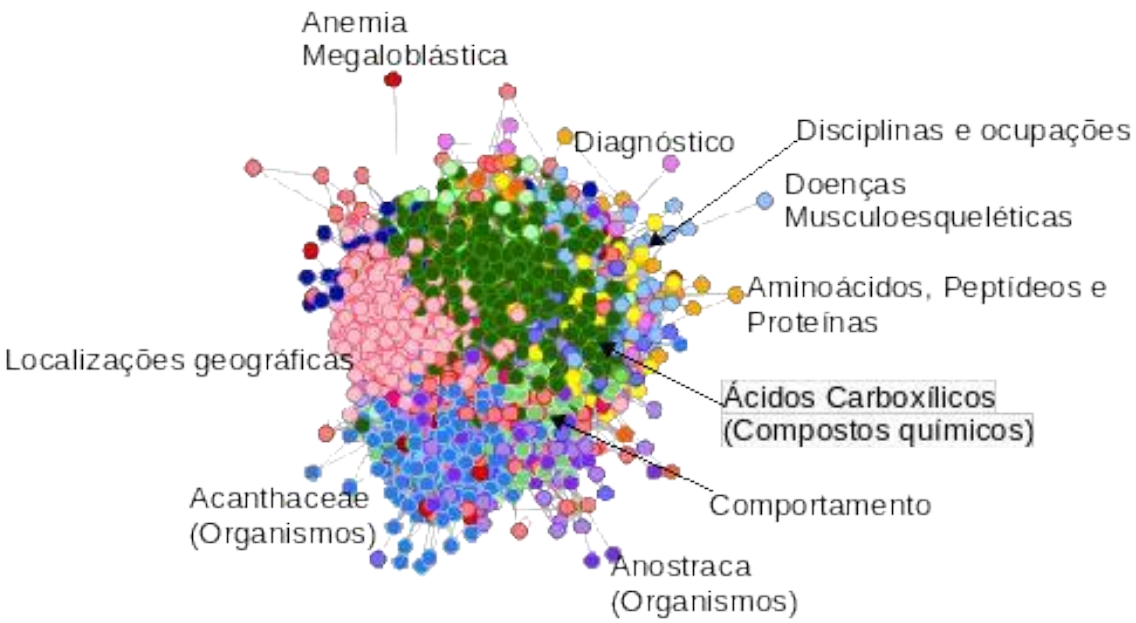


Figura 18. Parte da rede de termos controlados-consumidor mostrando 10 comunidades cujos vértices têm grau maior que 50. Versão interativa em: http://bit.ly/2McpTuG.

O modelo de rede proposto neste estudo possibilitou o mapeamento das interações entre termos de diferentes categorias e origem, a exemplo da Figura 19 que apresenta a estrutura do tópico oftalmologia/visão. Observa-se que ocorrem interações entre os termos controlados (círculo) e do consumidor (quadrado). Outro aspecto é a proximidade, baseada na distância entre vértices, entre os termos categorizados como “doenças” e como “técnicas, diagnósticos e equipamentos”, o que pode denotar uma característica intrínseca do tópico porque se trata de uma especialidade na qual o uso de equipamentos é predominante.

Na sequência os termos foram mapeados, unificados e indexados. Os identificadores (CUI e DUI) dos termos dos vocabulários do UMLS foram mantidos e outros foram criados para os termos não indexados previamente. Para os sinônimos ainda não relacionados aos termos preferidos, o que ocorreu ao importar os “termos relacionados”, uma nova coleta na base UMLS foi realizada. Com a finalidade de melhorar a consistência do conjunto de termos foram excluídos manualmente os referentes a categorizações exclusivas da DBpedia, como “espécies descritas em”.

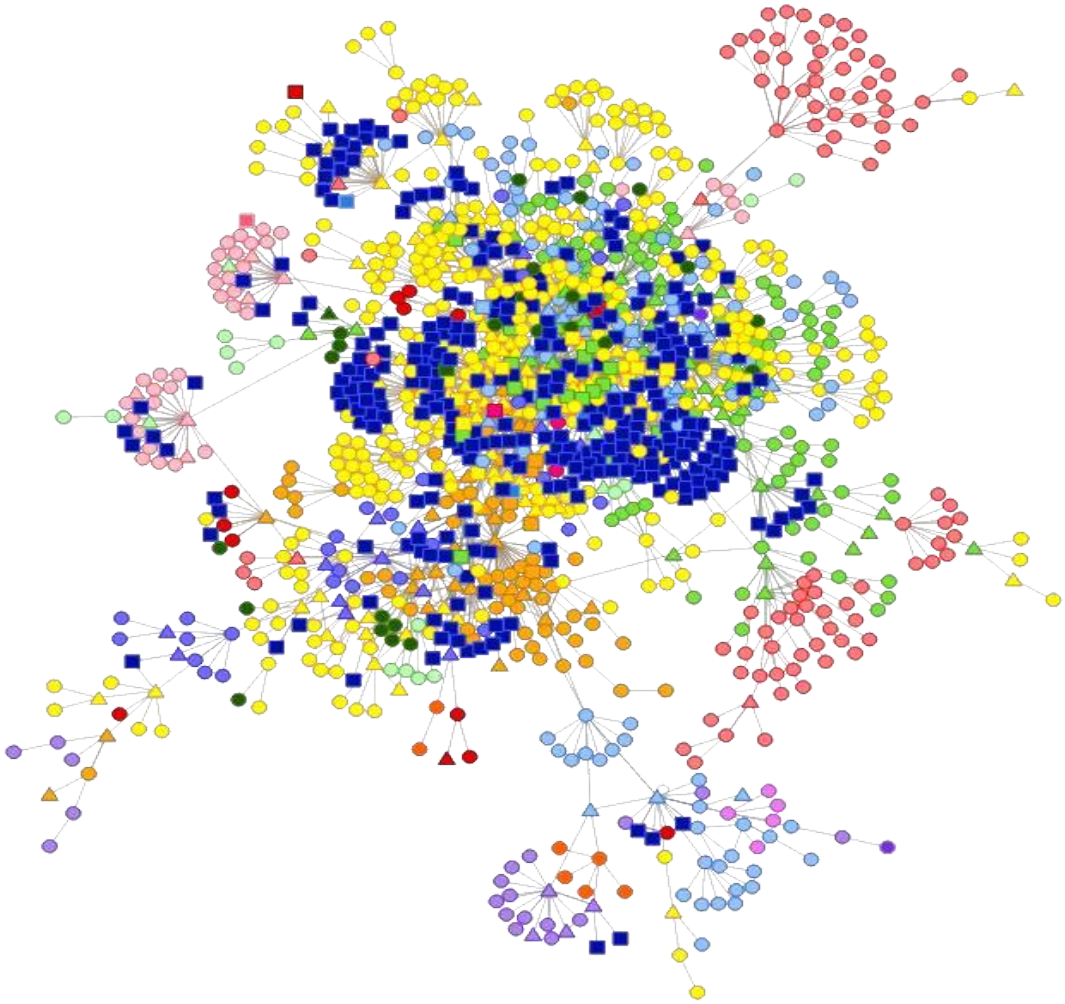


Figura 19. Detalhamento do agrupamento referente ao tópico oftalmologia/visão mostrando a estrutura das ligações entre termos provenientes de vocabulários controlado (círculo), do consumidor (quadrado) e de ambos (triângulo). As cores representam categorias: "doenças" (amarelo); "técnicas, diagnósticos e equipamentos" (azul escuro); "anatomia" (verde); "organismos" (rosa choque); "compostos químicos e drogas" (vermelho). Versão interativa em: http://bit.ly/2M77VcM.

Ao final da Fase 1 obteve-se um VSC composto por 146.956 termos, sendo que 31.439 são termos preferidos e 83.279 sinônimos referentes a eles (2,6 sinônimos/termo preferido). O aumento na quantidade de termos preferidos foi devido ao mapeamento de novos provenientes da base DBpedia. Ainda, há 32.238 termos a serem validados posteriormente como termo preferido ou sinônimo, mas que devem compor a estrutura de rede complexa proposta. A relação

sinônimo/termo preferido foi elevada de 1,6 (40.468/25.816) referente à coleta de dados dos vocabulários do UMLS/OAC-CHV para 2,5 (64.850/25.816) devido aos sinônimos recuperados da DBpedia.

A versão do VSC suportado por um modelo de rede complexa possibilitou a interação entre termos de categorias variadas, o que pode contribuir para a construção de aplicações mais poderosas, que explorem os aspectos referentes à estrutura semântica associada aos termos.

5.3 Identificação de termos e relações inéditos e validação (Objetivo 2)

Essa seção mostra os resultados referentes à estratégia 1 - recuperação de termos UMLS nos conteúdos não estruturados e estratégia 2 - extração de termos e relações semânticas inéditos, por meio da aplicação de técnicas de análise de textos.

5.3.1 Recuperação de termos UMLS em conteúdos não estruturados (Estratégia 1)

A ocorrência de termos presentes nos vocabulários do UMLS foi verificada nos conteúdos não estruturados coletados (NR, NC e PR). Essa estratégia possibilitou que termos coletados ou extraídos de vocabulários controlados não pertencentes à OAC-CHV, mas que ocorrem em conteúdos de domínio do consumidor fossem recuperados para compor o VSC. Considerou-se o pressuposto que, se um termo é utilizado em conteúdos disponíveis ao consumidor, deve constar de um vocabulário orientado a este devido à sua popularidade.

A quantidade de termos dos vocabulários do UMLS recuperados de cada conteúdo e o total de termos obtidos estão resumidos na Figura 20.

A Tabela 4 apresenta os 10 termos mais frequentes relacionados ao identificador e à sua categoria, que foi herdada do DeCS.

A Tabela 5 apresenta as categorias mais frequentes referentes aos termos do UMLS recuperados dos conteúdos não estruturados.

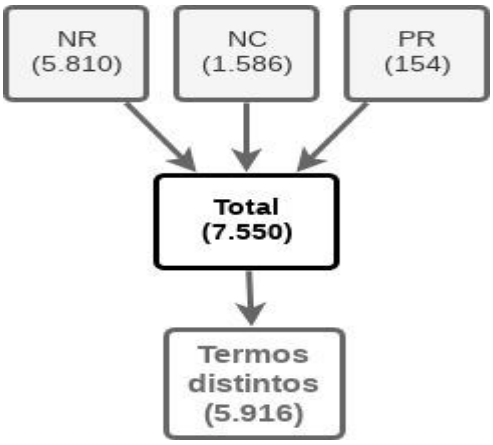


Figura 20. Quantidade de termos dos vocabulários UMLS recuperados de cada conteúdo e o total de termos distintos.

Tabela 4. Termos dos vocabulários do UMLS recuperados dos conteúdos não estruturados.

CUI	Termo	Categoria	Frequência
C0018684	Saúde	Assistência à saúde Homeopatia Saúde pública	103.814
C0006137	Brasil	Denominações geográficas	37.188
C0015671	Pai	Antropologia, educação, sociologia e fenômenos sociais Denominações de grupos Psiquiatria e psicologia	22.885
C0030705	Pacientes	Denominações de grupos Homeopatia	15.501
C0042776	Vírus	Organismos Saúde pública	13.494
C0011065	Morte	Doenças	13.431
C1261322	Exame	*	12.291
C0009421	Coma	Doenças	7.602
C0041703	EUA	Denominações geográficas	7.453
C0030551	Pais	Antropologia, educação, sociologia e fenômenos sociais Denominações de grupos Psiquiatria e psicologia	6.755

*não pertence ao DeCS

Tabela 5. Frequência de termos UMLS de acordo com a categoria.

Categoria	Frequência
Doenças	886
Compostos químicos e drogas	415
Organismos	360
Técnicas e equipamentos analíticos, diagnósticos e terapêuticos	233
Anatomia	219
Fenômenos e processos	198
Assistência à saúde	179
Psiquiatria e psicologia	142
Antropologia, educação, sociologia e fenômenos sociais	140
Denominações geográficas	123

Dentre todos os conteúdos foram recuperados 996 termos dos vocabulários do UMLS não mapeados na OAC-CHV. A Tabela 6 apresenta os 10 termos organizados segundo a frequência, em valores decrescentes.

Tabela 6. Termos do UMLS não mapeados na OAC-CHV.

CUI	Termo	Frequência
C0085563	Planos de saúde	4.723
C2936319	Informe	2.697
C1527258	Paralisia infantil	1.710
C0010226	Corte	1.353
C0345996	Milia	1.208
C3178908	SMS	1.054
C1615607	Vírus H1N1	747
C1998602	Refeições	718
C1962923	Dependentes	595
C1956345	Debates	573

5.3.2 Aplicação de técnicas de análise de textos (Estratégia 2)

A análise linguística é uma importante etapa para a execução do algoritmo RAKE. Exige a indicação das classes gramaticais como parâmetro de forma a constituir n-gramas que apresentem essa estrutura.

Com a finalidade de identificar termos cuja estrutura linguística fosse análoga aos dos vocabulários controlados, uma análise linguística foi realizada de forma a estabelecer um critério de seleção das classes gramaticais.

A Figura 21 apresenta as categorias mais frequentes quando os termos do UMLS são analisados. As categorias gramaticais mais frequentes são “substantivo” (noun), “adjetivo” (adj), além de elementos de junção como preposições, artigos e pronomes (adp, det).

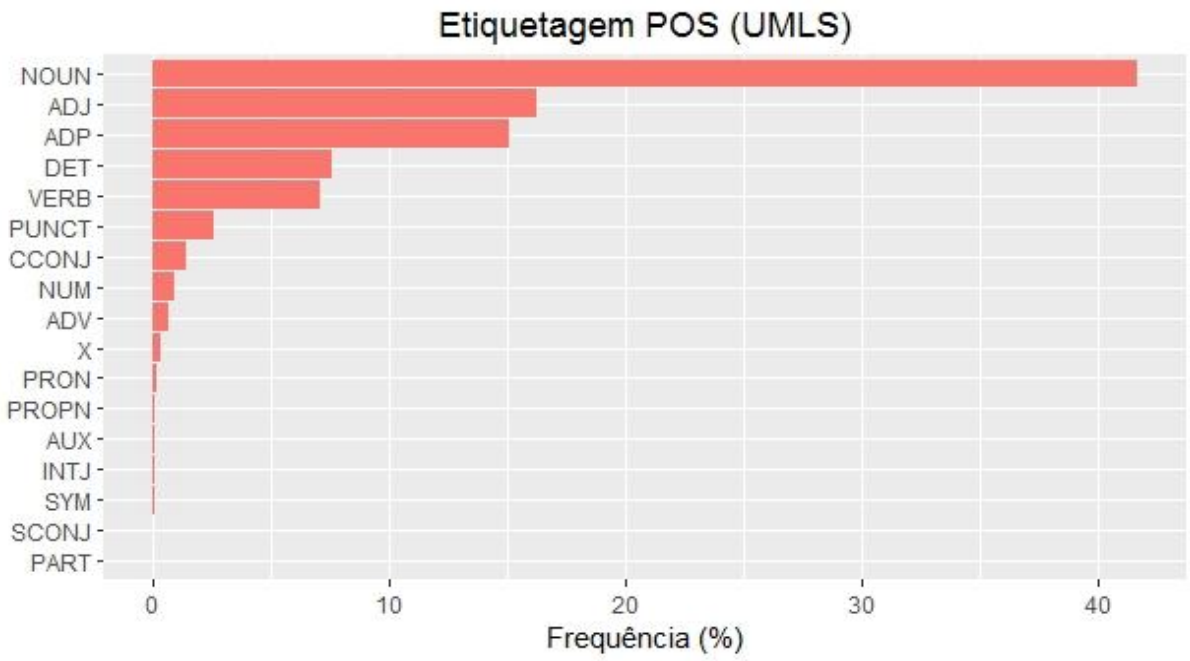


Figura 21. Estrutura linguística dos termos UMLS em idioma português (universaldependencies.org/u/pos/).

No primeiro ensaio essas quatro classes gramaticais foram utilizadas como valores do parâmetro utilizado para filtrar os n-gramas. Na sequência, a análise mostrou que mesmo os termos melhor classificados não apresentavam uma estrutura capaz de reconhecê-los como termos candidatos.

A condução do segundo ensaio, no qual os valores foram “substantivo” (noun) e “adjetivo” (adj), promoveu a identificação de n-gramas melhor caracterizados como termo e análogos aos termos previamente caracterizados como pertencentes à área de saúde.

A análise linguística (etiquetagem POS) foi realizada em cada conteúdo não estruturado, considerando os filtros mencionados. As Figuras 22 a 24 apresentam a frequência percentual de cada classe gramatical e outros elementos que compõem os conteúdos. Os gráficos mostram que os conteúdos, apesar de sua estrutura diferente, são equiparáveis.

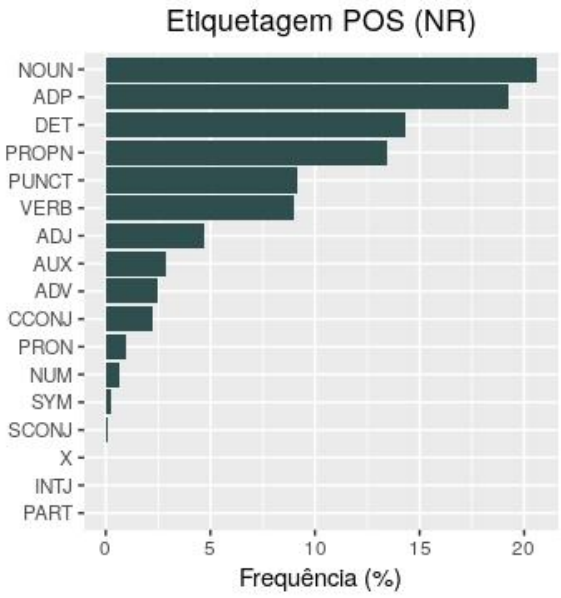


Figura 22. Etiquetagem POS mostrando a estrutura das notícias resumidas (NR).

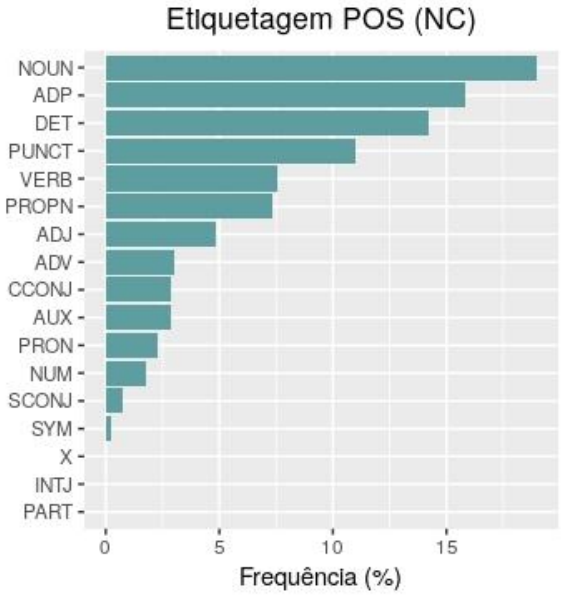


Figura 23. Etiquetagem POS mostrando a estrutura das notícias completas (NC).

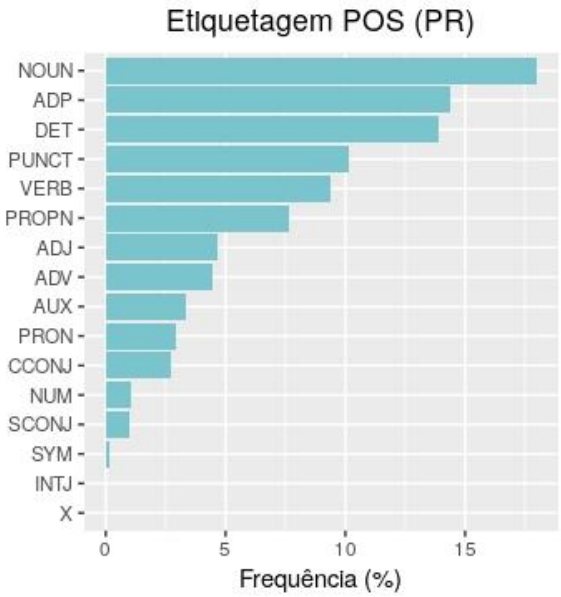


Figura 24. Etiquetagem POS mostrando a estrutura das perguntas e respostas (PR).

As Figuras 25 a 27 mostram o efeito da lematização em cada *token* referente a cada conteúdo. Os 30 *tokens* etiquetados como substantivos e adjetivos numericamente mais frequentes foram organizados.

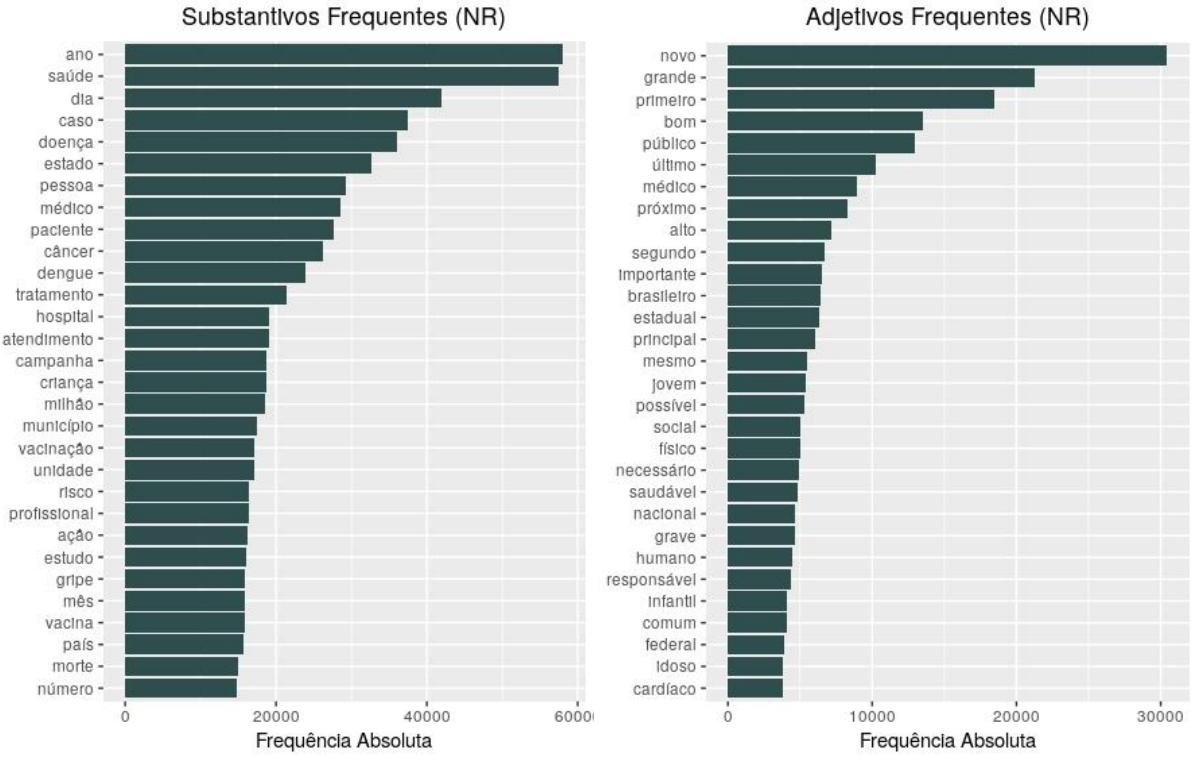


Figura 25. *Tokens* etiquetados como substantivos e adjetivos mais frequentes em notícias resumidas (NR).

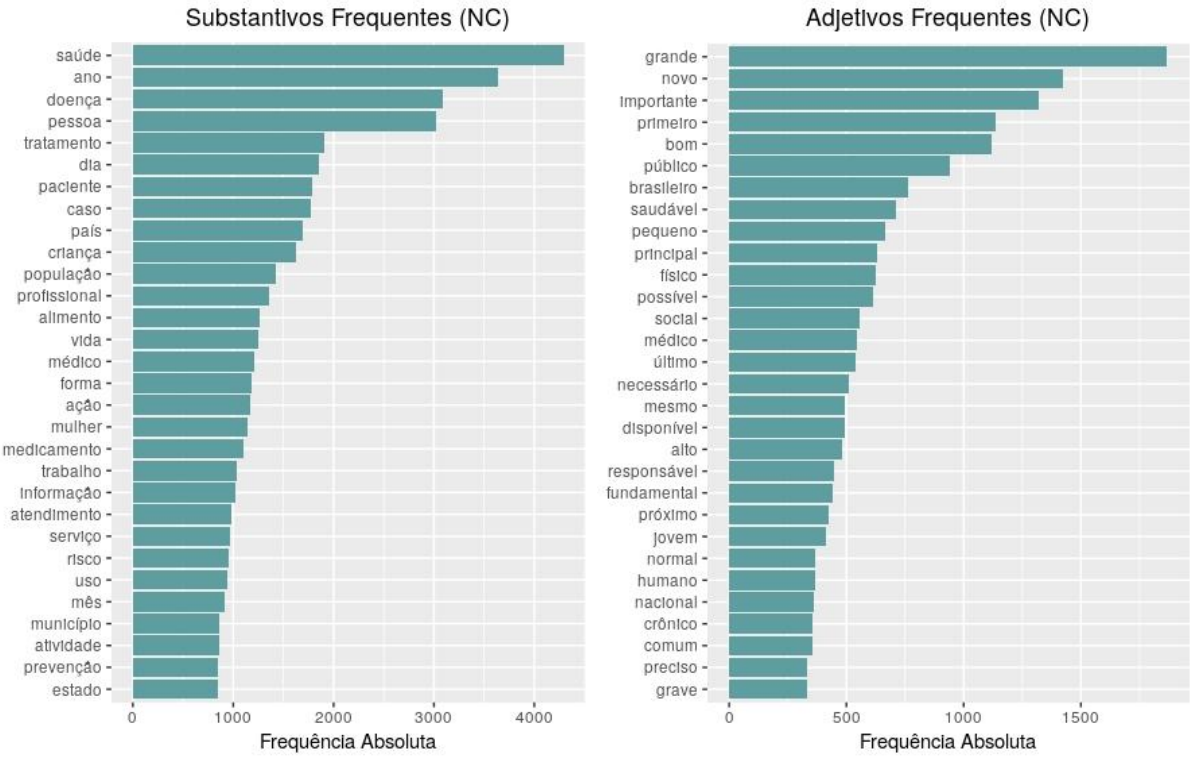


Figura 26. *Tokens* etiquetados como substantivos e adjetivos com maior número de ocorrências em notícias completas (NC).

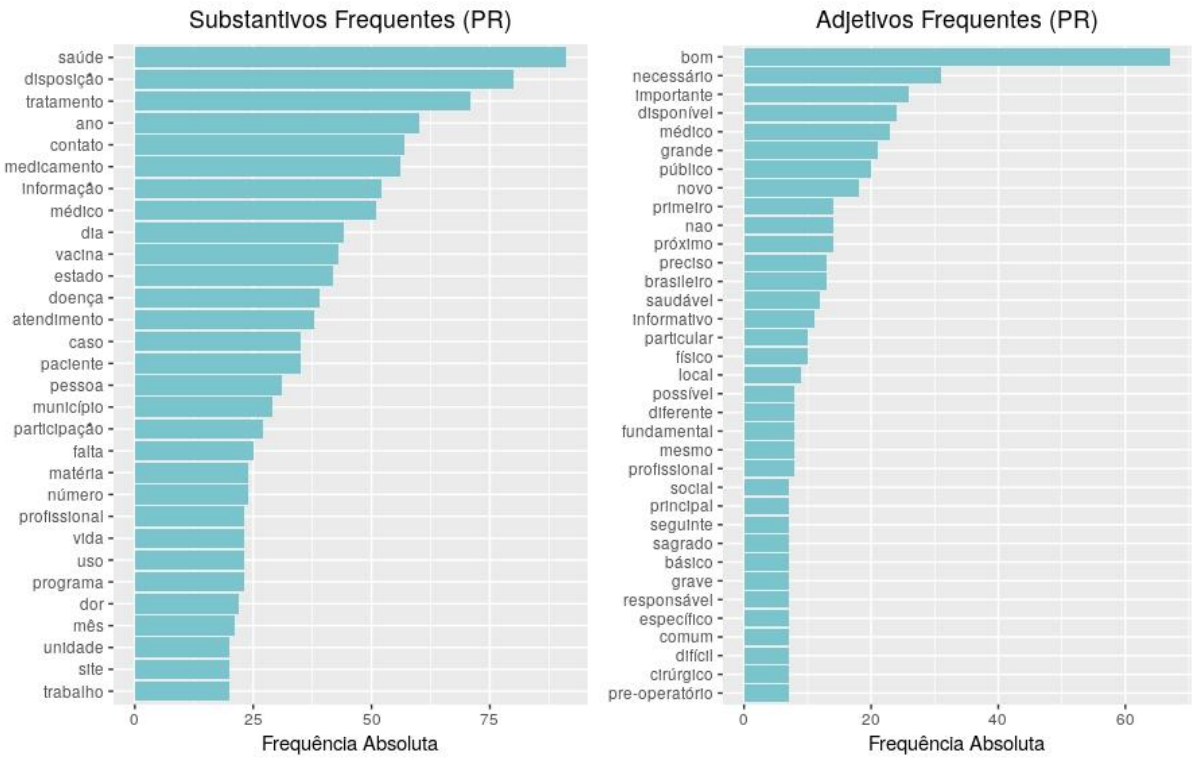


Figura 27. *Tokens* etiquetados como substantivos e adjetivos com maior número de ocorrências em perguntas e respostas (PR).

Observa-se que, mesmo os conteúdos versando essencialmente sobre temas de saúde, os *tokens* com maior número de ocorrências não apresentam especificidades da área.

A execução do algoritmo RAKE revelou uma lista de palavras-chave organizadas pelo índice próprio, considerando o efeito da lematização. A lista foi composta por 12.705 palavras-chave extraídas de NC, 77.583 de NR e 674 de PR considerando o índice com valor maior ou igual zero.

As Figuras 28 a 30 relacionam as 30 palavras-chave lematizadas, etiquetadas de acordo com as classes gramaticais substantivo e adjetivo, com maior índice RAKE obtidas para cada conteúdo.

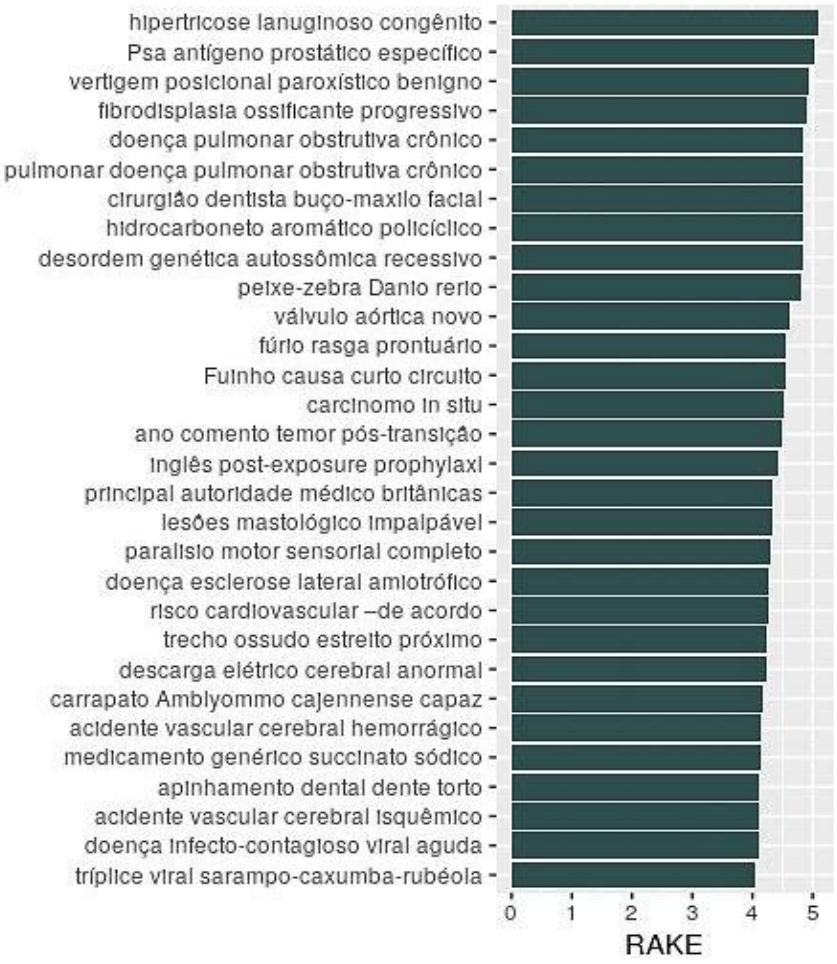


Figura 28. N-gramas lematizados com maior índice RAKE em notícias resumidas (NR).

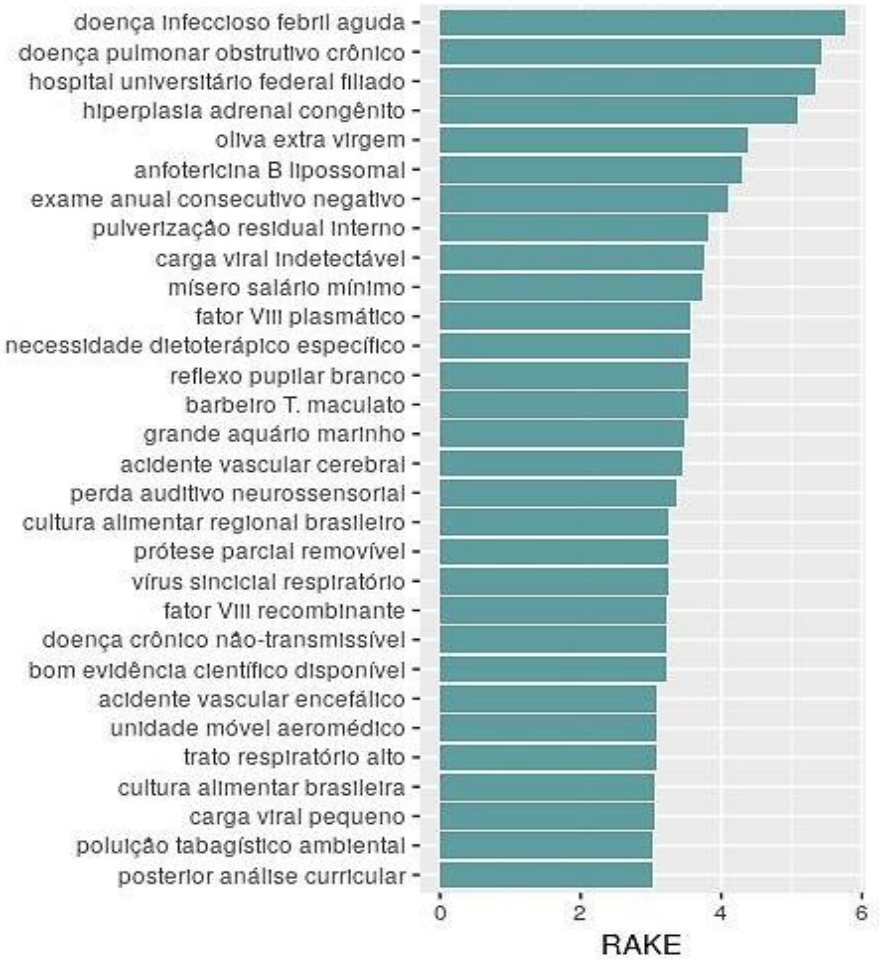


Figura 29. N-gramas lematizados com maior índice RAKE em notícias completas (NC).

Apesar do fato de que os *tokens* mais frequentes não serem específicos, o algoritmo foi capaz de identificar n-gramas específicos da área de saúde, com no máximo 5-gramas.

A distribuição do índice RAKE foi analisada por meio do gráfico apresentado na Figura 31. Todos os conteúdos retornaram um mesmo padrão de distribuição no qual a maior parte dos n-gramas com maior índice foram associados a menor ocorrência. De fato, como apresentado nas Figuras 25 a 27, *tokens* com maior número de ocorrências, normalmente, não foram específicos da área.

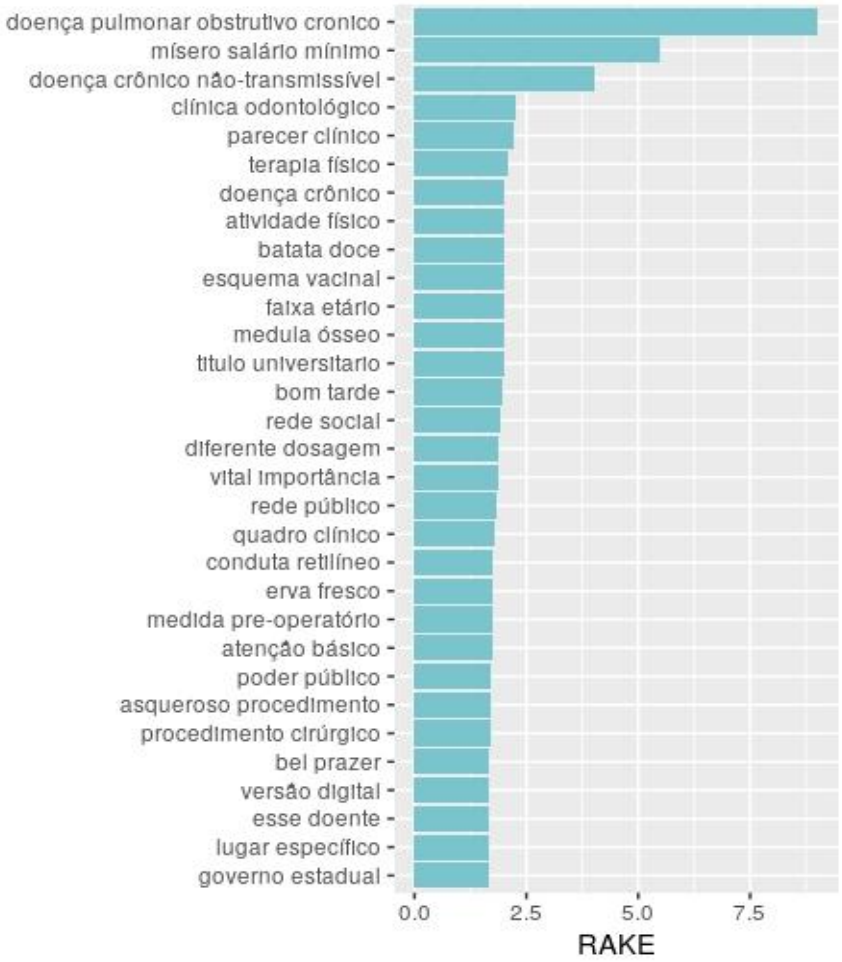


Figura 30. N-gramas lematizados com maior índice RAKE em perguntas e respostas (PR).

Um critério de corte foi então estabelecido para fins de compor uma lista de termos candidatos inéditos. Os n-gramas com índice RAKE maior ou igual a 1,5 foram eleitos para a etapa de validação. Esse valor foi estabelecido de forma empírica por meio de inspeção visual. Esses n-gramas foram considerados potencialmente mais relevantes porque, conforme indicado na Figura 31, ocupam a região do gráfico com maior índice RAKE, associado a menor frequência e, consequentemente, a *tokens* menos comuns. Observou-se que para valores menores que 1,5 os termos não foram específicos da área de saúde.

Uma lista de 9.674 n-gramas foi consolidada utilizando o ponto de corte do índice RAKE.

Os 30 n-gramas melhor classificados foram organizados no gráfico da Figura 32. Observa-se cerca de 22/30 (73%) n-gramas potenciais termos candidatos ao domínio da saúde.

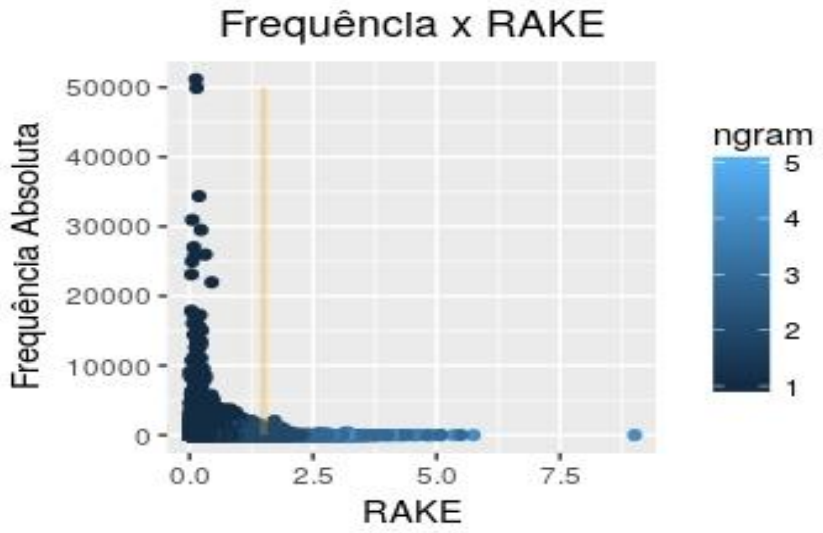


Figura 31. Distribuição do índice RAKE em relação ao número de ocorrências dos n-gramas, com destaque para o ponto de corte (1,5).

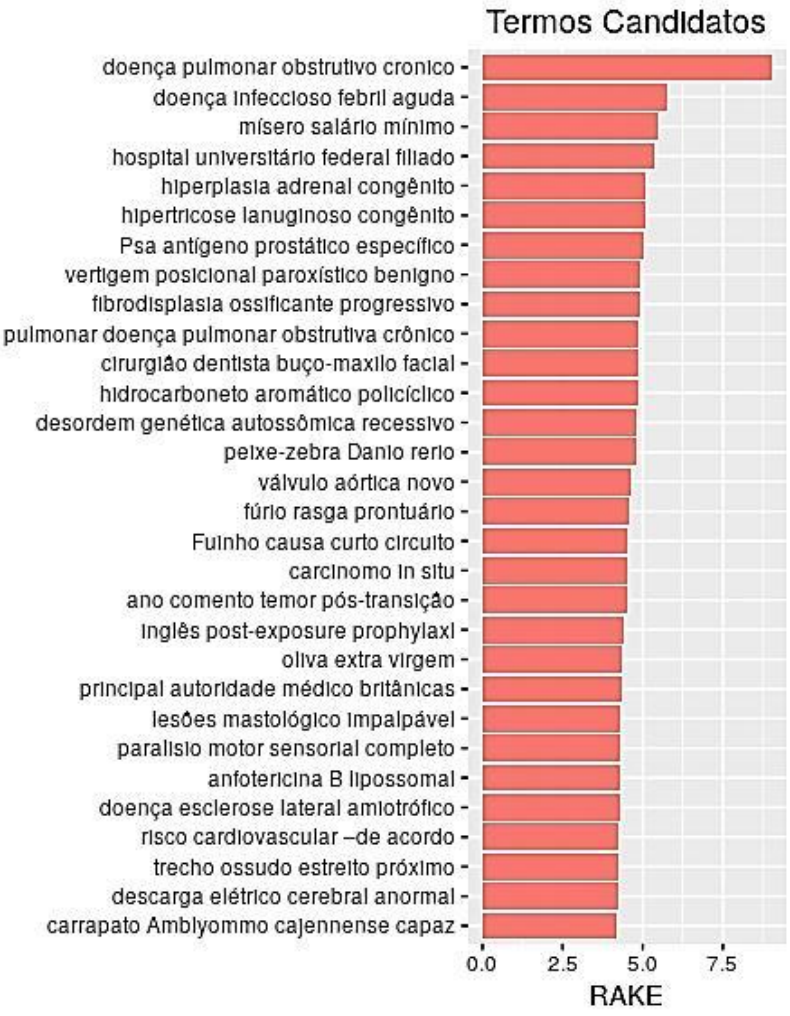


Figura 32. N-gramas lematizados com maior índice RAKE.

Para propiciar uma melhor análise dos termos identificados a Tabela 7 mostra 33 n-gramas com o valor de corte (1,5) estabelecido para o índice RAKE. Nota-se que 64% dos n-gramas são potenciais termos candidatos.

Tabela 7. N-gramas lematizados com índice RAKE = 1,5.

n	n-gramas	Frequência	n	n-gramas	Frequência
1	figura paterno	5	18	enzimo alfa-amilase	2
2	hipertensão arterial	4	19	enzimo aromatase	2
3	Psa	3	20	estriado ventral	2
4	cova raso	3	21	estrito observância	2
5	acanthamoeba keratitis	3	22	fendo palatina	2
6	braço esquerdo	3	23	fútbol clubado	2
7	levotiroxina sódico	3	24	hora-aula ministrado	2
8	processo seletivo	3	25	lamivudina 300mg	2
9	propriedade antioxidante	3	26	lâmpada fluorescente	2
10	Tamanho irresponsabilidade	3	27	nexo causal	2
11	Campanho	2	28	perfurador pneumático	2
12	cefa orgástico	2	29	pernilongo doméstico	2
13	cerebral	2	30	picanha invertida	2
14	Chaga	2	31	pneumonio bacteriana	2
15	delegação ministerial	2	32	puericultura correlato	2
16	Dermatite esfoliativo	2	33	unho engrossamento	2
17	dermo abrasão	2			

A coocorrência entre os n-gramas e os termos dos vocabulários do UMLS resultaram na identificação de relações semânticas, inicialmente do tipo “relacionado”. A Figura 33 mostra os n-gramas mais importantes segundo o índice RAKE, que coocorrem com o termo preferido “Poliomielite”. Dentre o conjunto de n-gramas há “polimielite” e “paralisia infantil”, que são sinônimos do termo indicado. O mesmo ocorre com o par “Puberdade Precoce-amadurecimento sexual precoce”.

As Figuras 34 e 35 também possibilitaram a identificação de pares termos

preferidos-sinônimo, a saber, “Traço falciforme - doença falciforme”, “Ataque isquêmico transitório - acidente vascular cerebral”. É válido citar que, na Figura 35, o par “Prótese parcial removível - ponte móvel” também ocorre. O valor de corte do número de n-gramas não possibilitou sua visualização completa. Observou-se a composição de outros pares “termo preferido-sinônimo”: “eritema multiforme - Síndrome de Stevens-Johnson” e “doença autoimune crônica - Sarcoidose”.

Além da associação de n-gramas e potenciais sinônimos, o agrupamento por meio de termos preferidos também possibilita a formação de tópicos, representados nas Figuras 33 a 35. A organização dos termos em tópicos é útil para suportar aplicações baseadas, por exemplo, na caracterização de assuntos específicos.

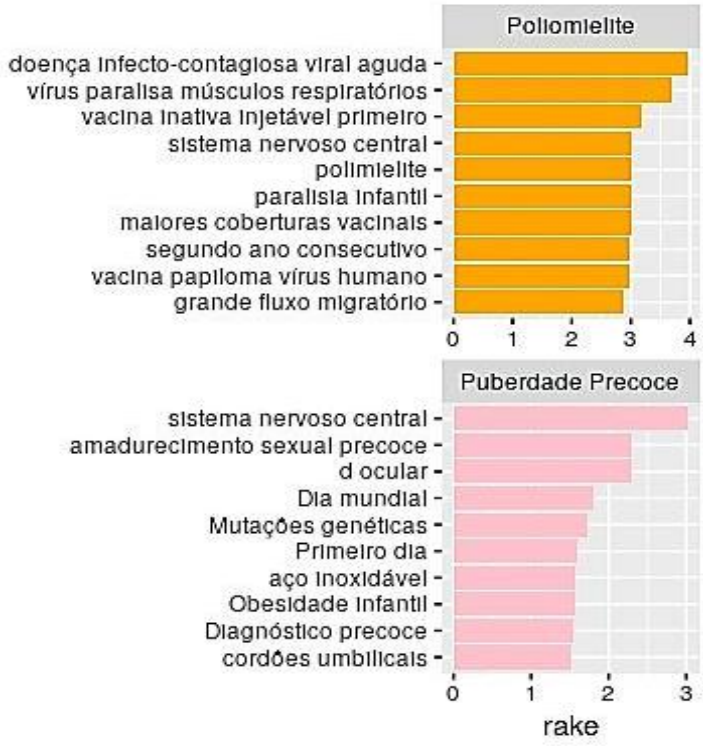


Figura 33. N-gramas associados por coocorrência aos termos preferidos do UMLS “Poliomielite” e “Puberdade Precoce”.

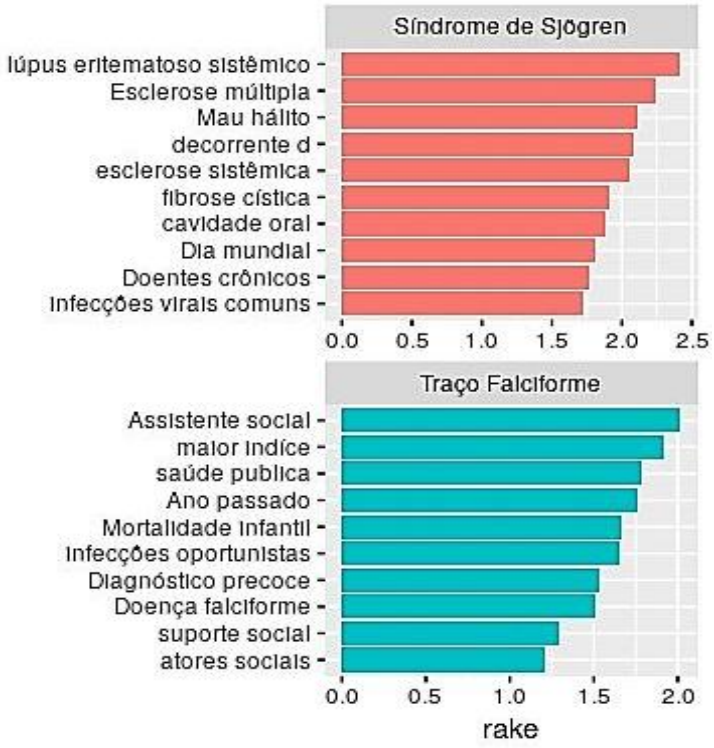


Figura 34. N-gramas associados por coocorrência aos termos preferidos do UMLS "Síndrome de Sjogren" e "Traço falciforme".

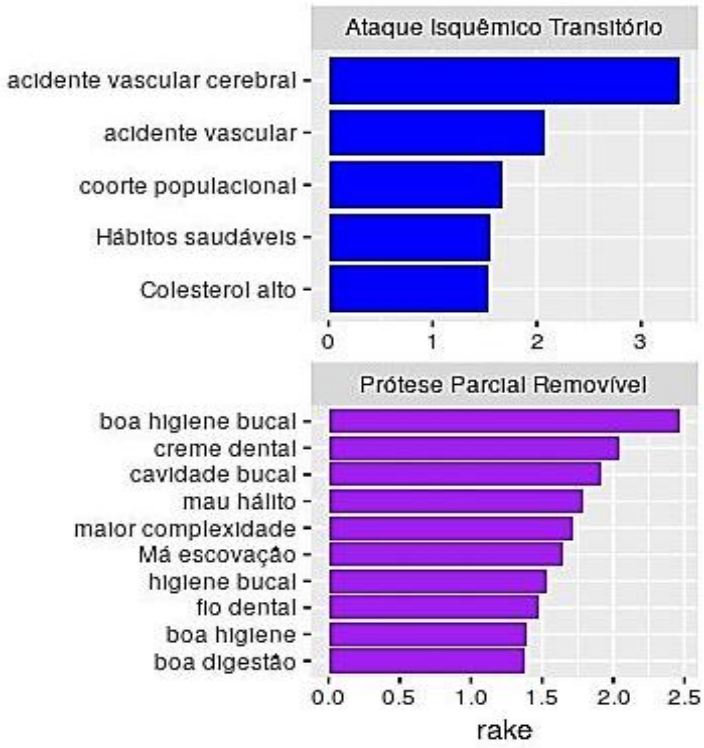


Figura 35. N-gramas associados por coocorrência aos termos preferidos do UMLS "Ataque isquêmico transitório" e "Prótese parcial removível".

A identificação de pares “termo preferido-sinônimo” foi importante para que no processo de validação humana este seja um item a ser avaliado.

A Figura 36 apresenta a associação por coocorrência entre os n-gramas e os termos dos vocabulários do UMLS, suportados por um modelo de rede.

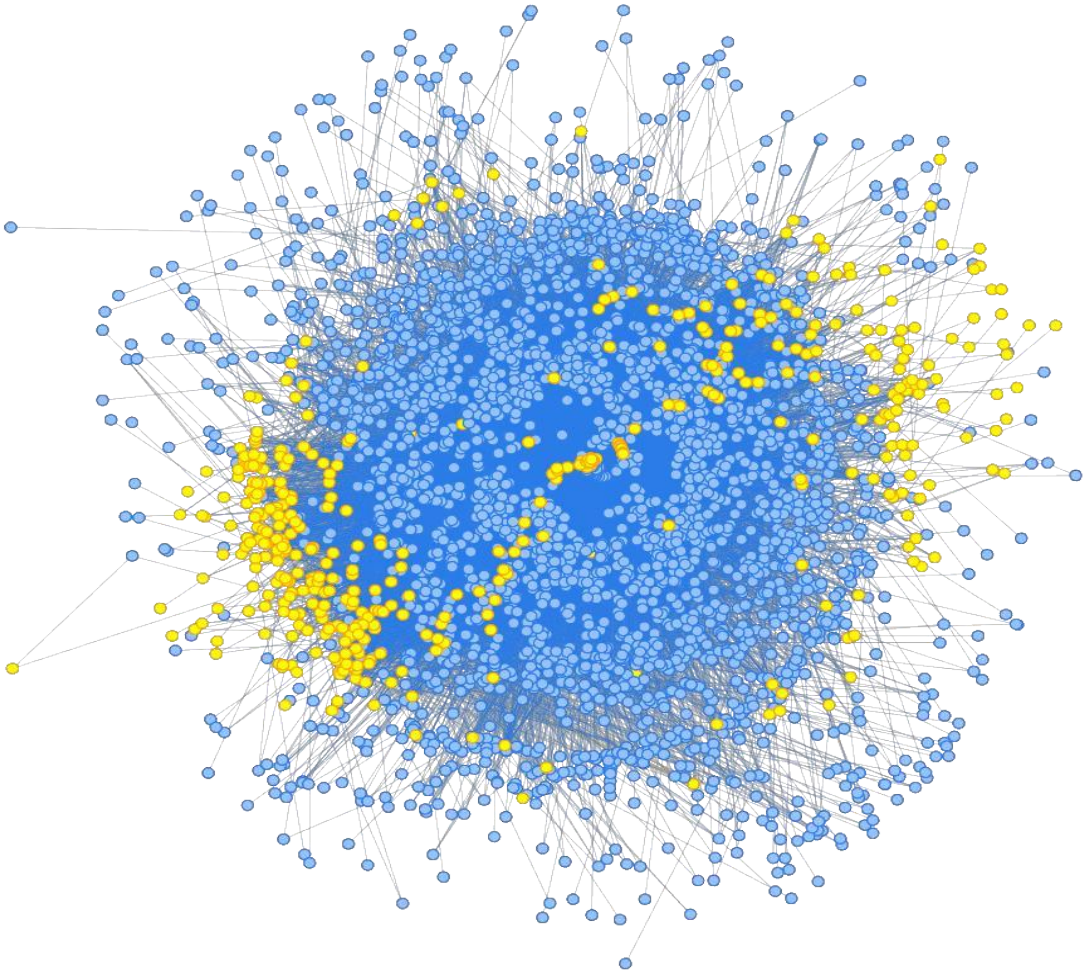


Figura 36. Modelo de rede complexa da associação por coocorrência entre n-gramas (azul) e termos do UMLS (amarelo). Versão interativa em https://bit.ly/2QHmdlO.

A Figura 37 mostra em detalhe uma região destacada de forma a possibilitar a identificação dos elementos.

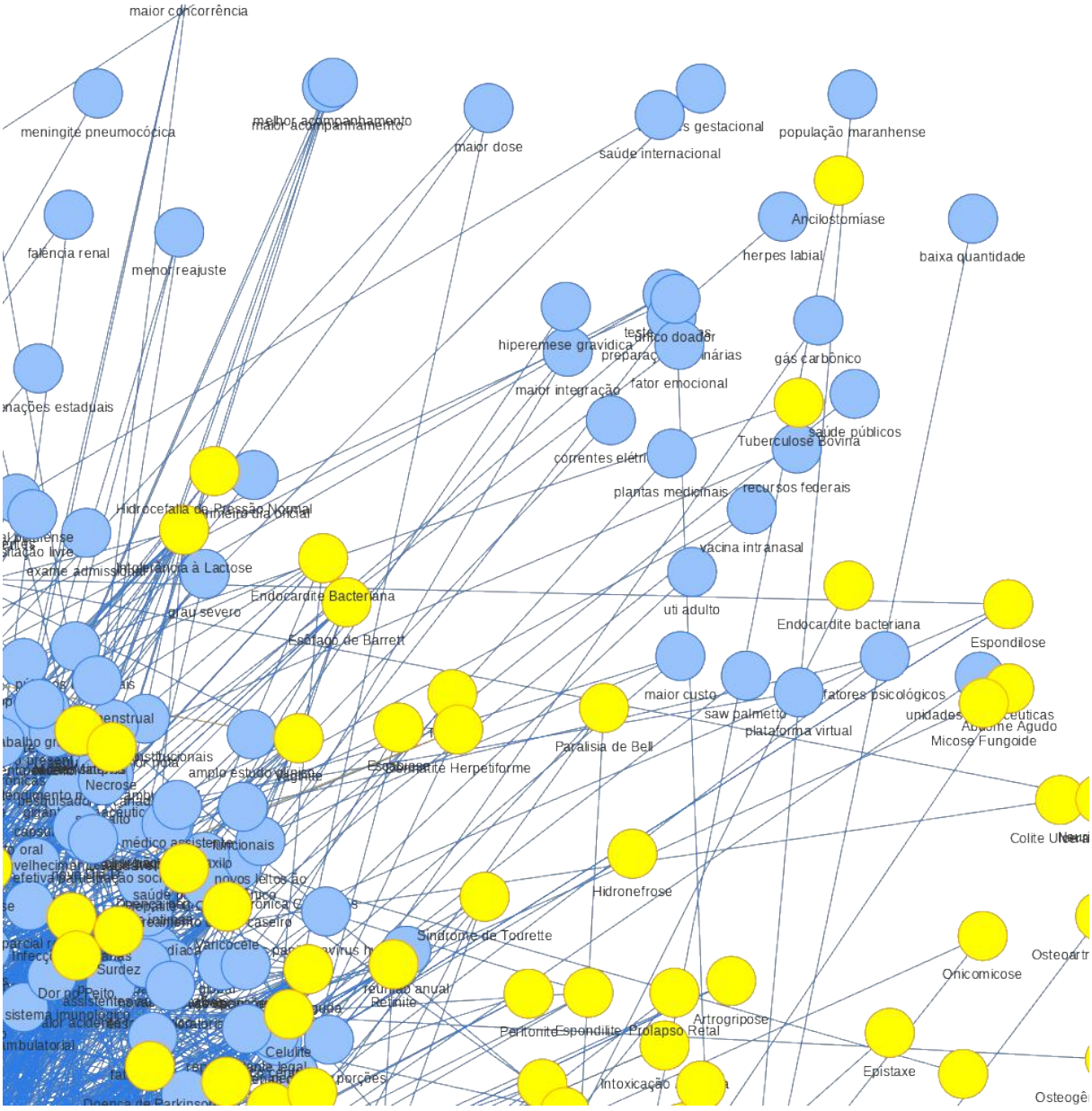


Figura 37. Detalhe evidenciando os vértices e arestas entre os n-gramas (azul) e termos dos vocabulários do UMLS (amarelo). Versão em alta resolução em https://bit.ly/2Psxl9l.

5.3.3 Validação humana

Foram disponibilizados 9.428 termos por meio da interface web. Cerca de 96% dos termos foram avaliados, no início do processo, por 79 profissionais da área de saúde e informática em saúde. A maior parte dos termos (87%) foi avaliada por 28 profissionais. A Tabela 8 apresenta a quantidade de termos avaliados segundo a classificação proposta.

Tabela 8. Síntese das respostas dos avaliadores.

Classificação	# termos	%
É saúde	3.821	42,2
Pode ser (saúde)	2.215	24,4
Não é saúde	1.958	21,6
Não tem sentido	872	9,6
Não sei	199	2,2
Total	**9.065**	**100**

Um segundo processo de avaliação foi conduzido. Os termos ainda não avaliados e os classificados como "não sei" foram reavaliados por uma profissional de saúde, que considerou apenas as opções "é saúde", "não é saúde", "pode ser saúde" e “não tem sentido”.

A necessidade subjacente a este processo é a decisão de importar os termos para o VSC. Para fins de validação conduziu-se uma releitura das respostas na qual os termos classificados pelos avaliadores como "é saúde" e "pode ser saúde" foram etiquetados como “validados” e os termos classificados como "não é saúde" e "não tem sentido", como “não validados”.

A mesma escala simplificada foi utilizada na avaliação da pesquisa principal de acordo com o protocolo estabelecido.

Análise estatística

A taxa de concordância entre as duas avaliações, considerando a validação dos termos, resultou em 78,7%.

A comparação entre a classificação dos termos pelos avaliadores (Figura 38) mostrou que ambos os avaliadores têm alta concordância ao validar termos classificados como “saúde”. Em menor grau o mesmo ocorre para as classificações “não é saúde”, e “pode ser saúde”. Neste estudo estabeleceu-se como padrão de termos de domínio da saúde os vocabulários controlados. A validação foi orientada segundo essa perspectiva. Uma possibilidade é que o resultado da validação e a diferença apresentada no gráfico foram decorrentes do desconhecimento de que mesmo termos comuns, como “Informe”, podem compor um vocabulário de saúde. O mesmo pode ser considerado na classificação de “termo inválido”, além do fato de que, ao desconsiderar alguns termos, os avaliadores não verificaram que os mesmos poderiam ser corrigidos, mesmo tendo sido orientados para esse fato.

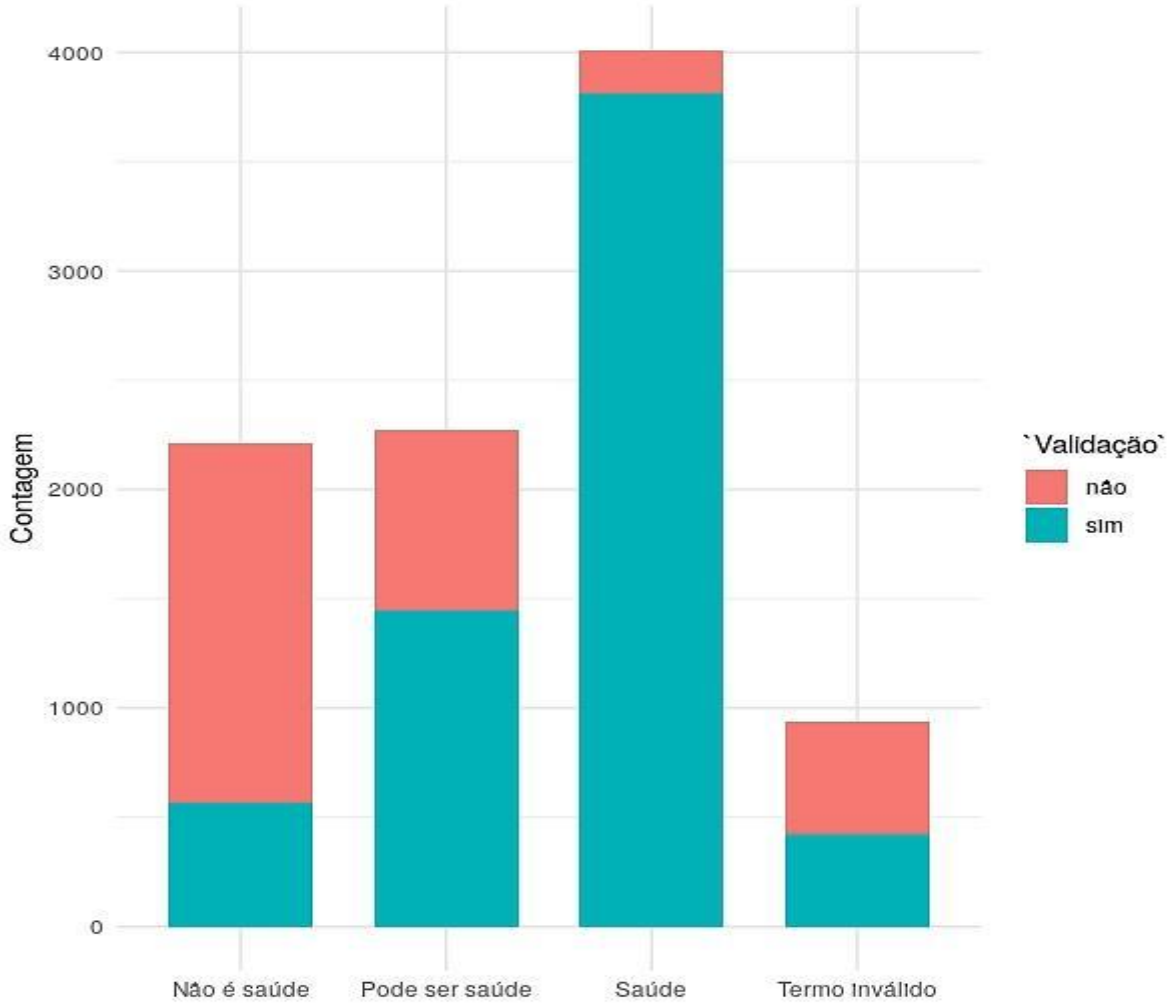


Figura 38. Comparação entre as avalições considerando a escala de validação.

O coeficiente de confiabilidade Kappa Cohen para dois avaliadores resultou em 0,54 (nível de confiança = 95%), classificado como "moderado".[107]

Configurou-se, neste caso, a situação em que a taxa de concordância é alta, mas apresenta baixo coeficiente de confiabilidade kappa, devido ao desequilíbrio entre as avaliações. Uma alternativa para essa situação é o coeficiente de confiabilidade de Gwet AC1 [108, 109], o qual por não considerar o pressuposto de independência entre os observadores pode ser usado em mais contextos que Kappa. O coeficiente Gwet AC1 resultou em 0,630±0,008, com nível de confiança 95%, qualificado como "bom".[110]

O coeficiente de viés sistemático entre dois avaliadores resultou em 0,491, com qui-quadrado de 0,643. Um viés sistemático entre avaliadores pode ser assumido quando a razão se desvia substancialmente de 0,5, o que não ocorreu

nesta análise.

A validação resultou em 6.245 termos admissíveis ao VSC, o que corresponde a 66,24% dos termos disponíveis. A quantidade de termos classificados como "sem sentido" ou "não é saúde" foi 3.188.

5.3.4 Análise do algoritmo RAKE

As medidas básicas como acurácia, sensibilidade e especificidade, assim como seus relacionamentos expressos pela curva ROC, foram realizadas para comparação com outros estudos. Entretanto, outras medidas mais adequadas ao contexto de recuperação da informação, como precisão, revocação e *F-measure*, também foram realizadas de forma a contribuir com a análise do algoritmo.

Outro aspecto importante é o fato de que, ao contrário de um classificador automático, no qual a matriz de confusão identifica os verdadeiros positivos e verdadeiros negativos, para o contexto de recuperação da informação não há como identificar os "termos negativos" ou "não termos". O resultado é que o valor dos verdadeiros negativos é sempre zero na matriz de confusão. Esse fator gera um desequilíbrio que determina um menor valor para medidas como área sob a curva ROC (AUC-ROC).

Medidas referentes ao desempenho do algoritmo RAKE na tarefa de identificar termo estão sumarizadas na Tabela 9. Os valores foram arredondados. O “rank normalizado” indica os limiares em cada quartil normalizado. As medidas são referentes a esses limiares.

Tabela 9. Avaliação de desempenho do algoritmo RAKE.

Medidas	Mínimo	1º. Quartil	Mediana	Média	3º. Quartil	Máximo
Rank normalizado	0,000	0,250	0,500	0,500	0,750	1,000
Acurácia	0,338	0,478	0,537	0,531	0,603	0,662
Especificidade	0,000	0,283	0,555	0,546	0,838	1,000
Sensibilidade	0,000	0,295	0,528	0,523	0,767	1,000
Precisão	0,662	0,677	0,700	0,726	0,781	1,000
MCC*	0,004	0,049	0,075	0,079	0,110	0,148
F-score	0,000	0,428	0,602	0,547	0,719	0,797

* coeficiente de correlação de Matthews.

A acurácia indica a proporção correta em relação à predição e é mais

adequada para conjunto de dados com classificações balanceadas. Neste estudo, devido ao fato de que cerca de que aproximadamente 70% dos valores pertencerem à classe que valida o termo, o coeficiente de correlação de Matthews (MCC) é mais robusto porque considera o equilíbrio entre as quatro categorias de uma matriz de confusão (verdadeiros positivos, verdadeiros negativos, falsos positivos, falsos negativos). O coeficiente varia entre +1 e -1; o valor +1 representa uma predição perfeita.

A Figura 39 apresenta as curvas ROC e precisão-revocação (PR) para o conjunto de termos avaliados. A curva ROC é apropriada quando as observações são balanceadas entre cada classe, enquanto a de precisão-revocação é apropriada para conjuntos de dados desbalanceados.[111]

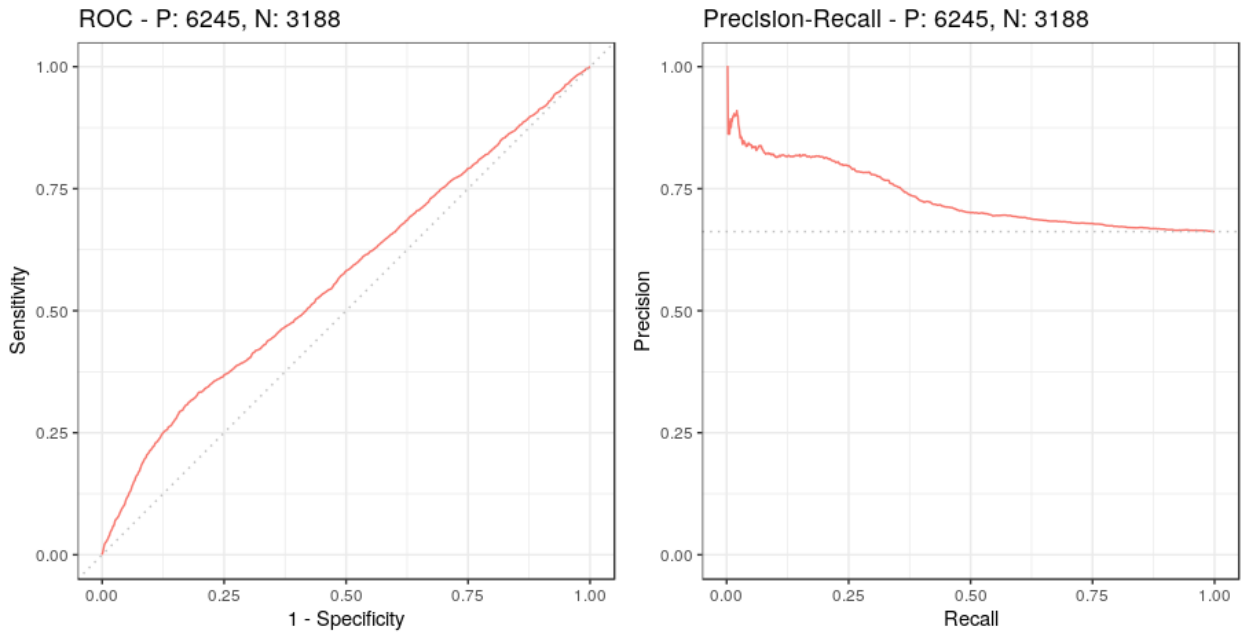


Figura 39. Curvas ROC e precisão-revocação (PR).

A área sob a curva ROC resultou em **0,569**, um valor considerado moderado. Neste estudo este valor indica a probabilidade do algoritmo RAKE de identificar os termos na área de saúde. O desempenho de RAKE poderia ser melhor se considerássemos apenas a capacidade de identificar termos em geral.

A curva PR, entretanto, descreve a capacidade do algoritmo RAKE em prever a classe positiva, ou seja, os verdadeiros termos. O valor resultante foi **0,732.** Este

valor indica a probabilidade do algoritmo RAKE de identificar termos válidos. A análise visual da curva PR mostra que essa é a probabilidade mínima de predição de termos válidos e que, para os primeiros 1.000 termos com maior índice RAKE (10%), pode chegar a aproximadamente 0,900.

5.4 Aplicação computacional (Objetivo 3)

A Fase 3 tratou da viabilização da tecnologia RDF para armazenamento do conteúdo e relações do VSC e do desenvolvimento de uma aplicação computacional para acesso aos dados. Para isso foi necessário organizar os conteúdos resultantes das Fases 1 e 2.

5.4.1 Composição do VSC

A partir da rede complexa de termos controlados-consumidor, construída na Fase 1, uma base de dados foi organizada composta por 146.956 termos, além de identificadores, relações semânticas, quando mapeadas, e sua origem. Expressões regulares foram aplicadas para unificar termos de forma a remover duplicações.

Essa base foi complementada com os termos validados na Fase 2, pela qual foram extraídos 996 termos dos vocabulários do UMLS, ainda não mapeados no conteúdo da OAC-CHV, e identificados 6.245 termos candidatos validados, reduzidos a 5.362 termos após a correção.

Na Fase 2 a correção possibilitou mapear 112 n-gramas compostos por termos reconhecidos como sinônimos associados a termos preferidos, a exemplo de "apinhamento dental dente torto". Os termos e o relacionamento foram mapeados e inseridos no VSC.

Os termos obtidos na Fase 2 foram associados aos termos dos vocabulários controlados do UMLS por meio de coocorrência. Esse mapeamento foi registrado como "termo relacionado".

Uma base foi consolidada resultando em 150.995 termos caracterizados por suas relações semânticas e origem, além da indexação única e identificadores UMLS, quando disponíveis. Há 66.992 conceitos ou potenciais conceitos,

identificados por termos preferidos, e 84.003 sinônimos.

5.4.2 Arquivo RDF

Para o armazenamento de conteúdo em modelo de dados RDF a ontologia SKOS[86] mostrou-se adequada a representação dos dados e suas relações, associada à ontologia PAV, que possibilitou o mapeamento de metadados como a origem do termo.

SKOS fornece três propriedades para anexar etiquetas a recursos conceituais: *skos: prefLabel*, *skos: altLabel* ou *skos: hiddenLabel*.[112]

O Quadro 3 esquematiza a estrutura de mapeamento de dados do VSC de acordo com as classes e propriedades propostas nas ontologias SKOS e PAV.

Quadro 3. Estrutura dos dados do VSC representados pela ontologia SKOS e PAV.

Dado	Tipo de propriedade	Propriedade	Subpropriedade
Conceito: (CUI/VSCID)		skos:concept	
inScheme			
DUI		dbo:meshId	
Termo preferido	Etiquetagem	rdfs:label	
Termo preferido	Etiquetagem	skos:prefLabel	
Sinônimo	Etiquetagem	skos:altLabel	
Sinônimo (variante)	Etiquetagem	skos:hiddenLabel	
Categoria	Relações semânticas	skos:semanticRelation	skos:broaderTransitive skos:narrowerTransitive
Relacionado	Relações semânticas	skos:semanticRelation	skos:related
Definição	Documentação	skos:note	skos:definition
Origem (Base de dados)	Proveniência	pav:importedFrom	
Origem (Processo)*	Proveniência	pav:derivedFrom	

* Processo que gerou o conteúdo obtido por refinamento ou modificação.

Os metadados do VSC também foram armazenados em triplas, por meio de propriedades referentes à data de criação do recurso, autoria, definição e outros.

O pacote rdflib [93] foi a base para a representação do VSC de acordo com o padrão RDF, baseado em triplas (sujeito, predicado e objeto). Para o sujeito foram criadas URIs, baseadas na utilização do CUI ou na criação de um identificador próprio do VSC (VSCID). A tarefa seguinte foi converter e armazenar os dados em várias serializações, a saber, rdfxml, nquads, turtle, ntriples e json-ld.

O VSC-RDF é composto por mais de um milhão de triplas. Para fins de visualização do conteúdo e estrutura, de acordo com as ontologias selecionadas, um exemplo foi criado (Figura 40) por meio do conceito identificado pelo CUI C0004763, etiquetado por "Esôfago de Barrett".

```
@base <http://telemedicina.unifesp.br/projeto/vsc> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix skos: <http://www.w3.org/2008/05/skos#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix pav: <http://purl.org/pav/> .
@prefix vsc_id: <.> .
@prefix vsc: <#> .
@prefix dbo: <http://pt.dbpedia.org/resource/> .

<#550053>
    skos:related vsc:C0004763 .

vsc:C0004763
    rdfs:label "Esôfago de Barrett" ;
    owl:sameAs dbo:Esôfago_de_Barrett, dbo:Syd_Barrett ;
    skos:Concept "esofago de barrett" ;
    skos:altLabel "Esofagite de Barrett", "Syd Barrett", "Síndrome de Barrett" ;
    skos:broaderTransitive vsc:C0012242, vsc:C0017163, vsc:C0266015 ;
    skos:hiddenLabel "esofagite de barrett", "sindrome de barrett", "syd barrett" ;
    skos:inScheme vsc_id:vsc ;
    skos:prefLabel "Esôfago de Barrett" ;
    skos:related vsc:C0034069, vsc:C0036864 .

vsc:C0014859
    skos:related vsc:C0004763 .

vsc:C0014876
    skos:related vsc:C0004763 .

vsc:C0017168
    skos:related vsc:C0004763 .

vsc:C0039082
    skos:related vsc:C0004763 .

vsc_id:vsc55756
    pav:importedFrom "MSHPOR" ;
    vsc:ID "Esôfago de Barrett" .

```

Figura 40. Exemplo do arquivo RDF em formato "turtle". Disponível em: http://bit.ly/31MXsKq.

O mesmo conceito foi usado para a construção do grafo RDF (Figura 41), que possibilita a observação das propriedades e objetos relacionados ao sujeito.

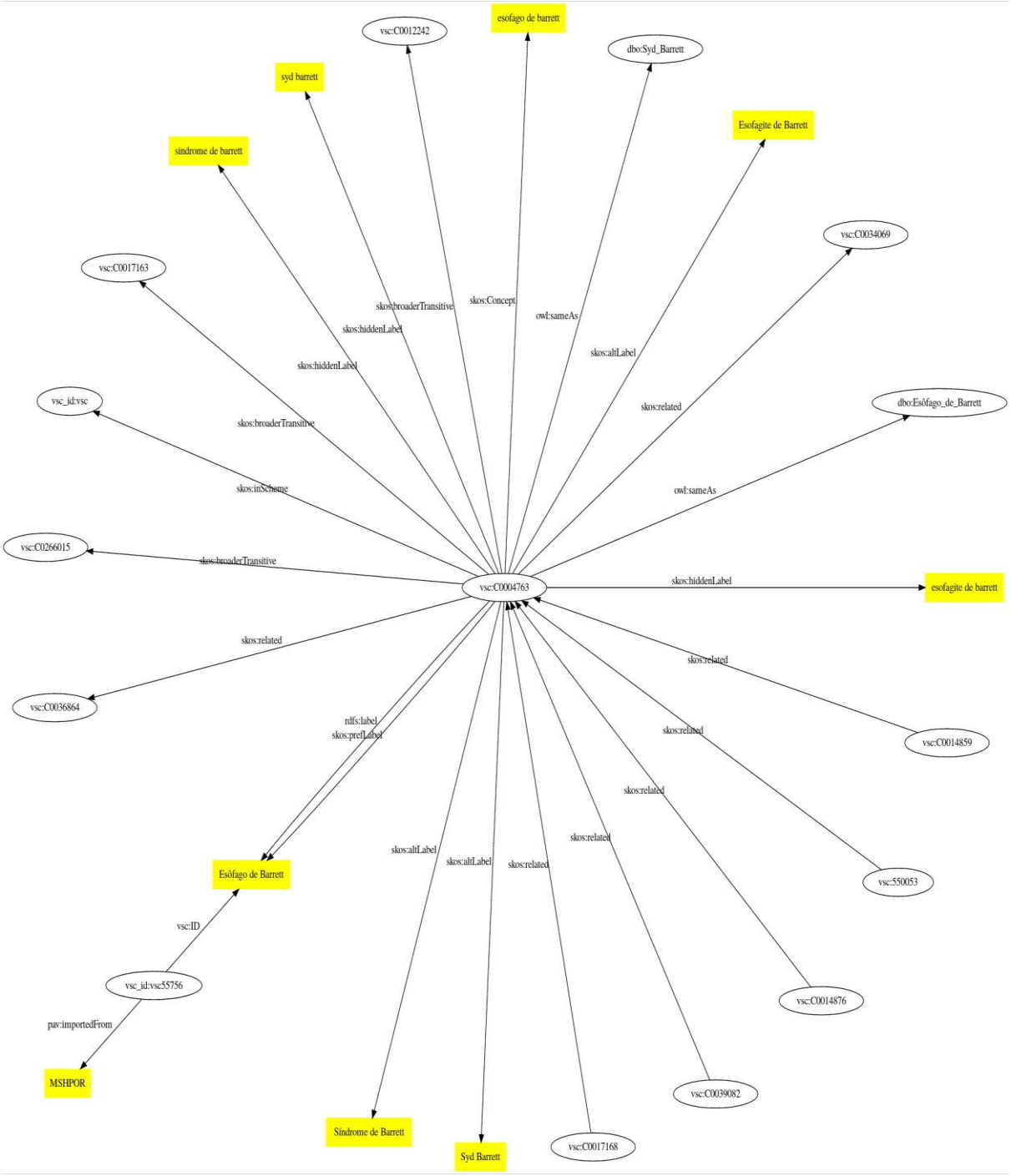


Figura 41. Exemplo de grafo RDF. Disponível em: http://bit.ly/2MbjnEo.

A linguagem de consulta SPARQL[94] possibilitou recuperar os dados armazenados no grafo RDF. O pacote rdflib suporta essa linguagem e apresenta os resultados em uma tabela, como mostra a Figura 42. A consulta retornou a categoria, cujo predicado é *skos:broaderTransitive*, associada ao termo preferido, por meio do predicado *skos:prefLabel* para um recurso específico, identificado pela

URI *vsc:C0004763*. O predicado *skos:broaderTransitive* é importante para o VSC porque associa o termo a um significado mais amplo, de forma indireta.

```
> sparql <-
+ 'PREFIX skos: <http://www.w3.org/2008/05/skos#>
+ PREFIX vsc: <http://telemedicina.unifesp.br/projeto/vsc#>
+ SELECT ?conceito ?categoria
+ WHERE {?termo skos:prefLabel ?conceito .
+ ?termo2 skos:prefLabel ?categoria .
+ ?termo skos:broaderTransitive ?termo2 .
+ FILTER (?termo = <http://telemedicina.unifesp.br/projeto/vsc#C0004763>) .
+ }'
> rdf_query(some_rdf, sparql)
# A tibble: 3 x 2
  conceito           categoria
  <chr>              <chr>
1 Esôfago de Barrett Doenças do Sistema Digestório
2 Esôfago de Barrett Gastroenterologia
3 Esôfago de Barrett Anormalidades do Sistema Digestório
>
```

Figura 42. Consulta SPARQL para recuperação da categoria e termo preferido referente ao conceito identificado por C0004763.

Consultas SPARQL podem recuperar todos os sinônimos armazenados por meio do predicado *skos:altLabel*, ou restringir a busca para um recurso específico, como exemplifica a consulta SPARQL na Figura 43.

```
> sparql <-
+ 'PREFIX skos: <http://www.w3.org/2008/05/skos#>
+ PREFIX vsc: <http://telemedicina.unifesp.br/projeto/vsc#>
+ SELECT ?Conceito ?Sinonimo
+ WHERE {?termo skos:prefLabel ?Conceito .
+ ?termo skos:altLabel ?Sinonimo .
+ FILTER (?termo = <http://telemedicina.unifesp.br/projeto/vsc#C0004763>) .
+ }'
> rdf_query(some_rdf, sparql)
# A tibble: 3 x 2
  Conceito           Sinonimo
  <chr>              <chr>
1 Esôfago de Barrett Syd Barrett
2 Esôfago de Barrett Esofagite de Barrett
3 Esôfago de Barrett Síndrome de Barrett
>
```

Figura 43. Consulta SPARQL para recuperação de sinônimos filtrados pelo conceito C0004763.

Outras consultas podem levar à recuperação de termos relacionados por meio do predicado *skos:related*, ou mesmo da origem do termo por meio dos predicados *pav:importedFrom* ou *pav:derivedFrom*.

5.4.3 Interface para acesso ao VSC

Uma interface web foi desenvolvida para propiciar a pesquisa de termos e recuperação da estrutura semântica mapeada (Figura 44). Os termos relacionados possibilitam a navegação entre outros termos apresentados. Os usuários poderão explorar o conteúdo do VSC e analisar a confiabilidade dos resultados de acordo com a proveniência do termo. A interface web foi disponibilizada para testes apenas para acesso interno no endereço 172.22.164.35:8081.

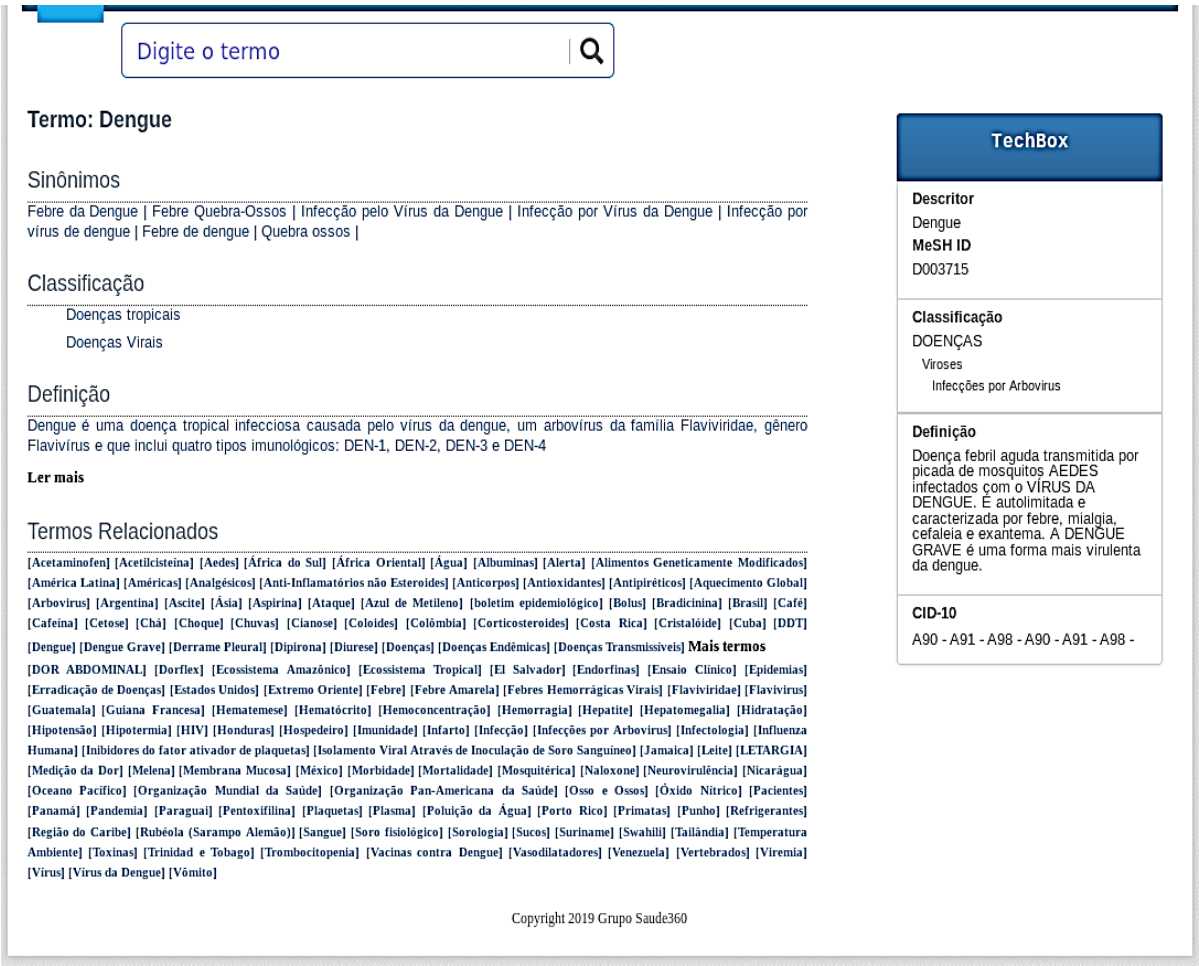
O acesso aos sinônimos, identificação e proveniência também está acessível para pesquisa (Figura 45).

A navegação interativa diretamente por meio de um grafo com 800 vértices (Figura 46), o limite de vértices renderizados na aplicação, foi disponibilizada para fins de exploração do conteúdo e estrutura de parte do VSC. Ao clicar em um vértice uma caixa com informações sobre o termo preferido apresenta sinônimos, classificação e termo relacionado. A origem do termo preferido está associada à cor do vértice.

Um aspecto importante na pesquisa sobre CHV é a confiabilidade dos termos. Neste estudo a proveniência do termo foi apresentada ao usuário para que este possa ter um parâmetro para avaliar o conteúdo apresentado.

A atualização dos vocabulários requer esforços computacionais e humanos. Uma forma para a contribuição dos usuários à ampliação do conteúdo e sua verificação é a disponibilização de uma interface colaborativa que propicie ao usuário propor termos inéditos ou mesmo propor correções aos termos presentes no conteúdo. Para promover essa finalidade uma interface foi desenvolvida (Figura 47) para que o usuário indique possíveis correções ou sugira novos termos preferidos ou sinônimos.

(a)



(b)

Figura 44. Interface para pesquisa de termos (a) e apresentação do resultado da pesquisa por termo e estrutura semântica (b).

Figura 45. Interface para pesquisa de termos preferidos e sinônimos.

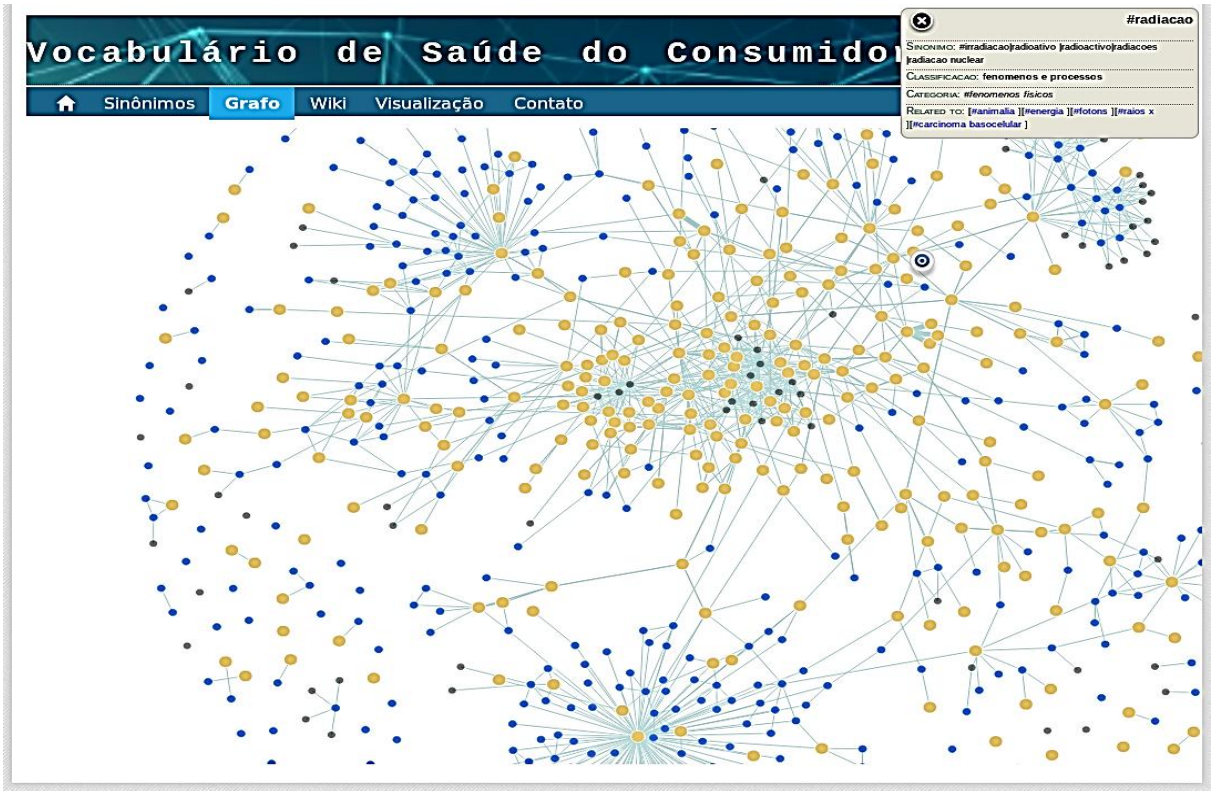


Figura 46. Interface para navegação no grafo. A origem dos vértices está associada à cor: vocabulários UMLS (azul), DBpedia (marron) e ambos (cinza).

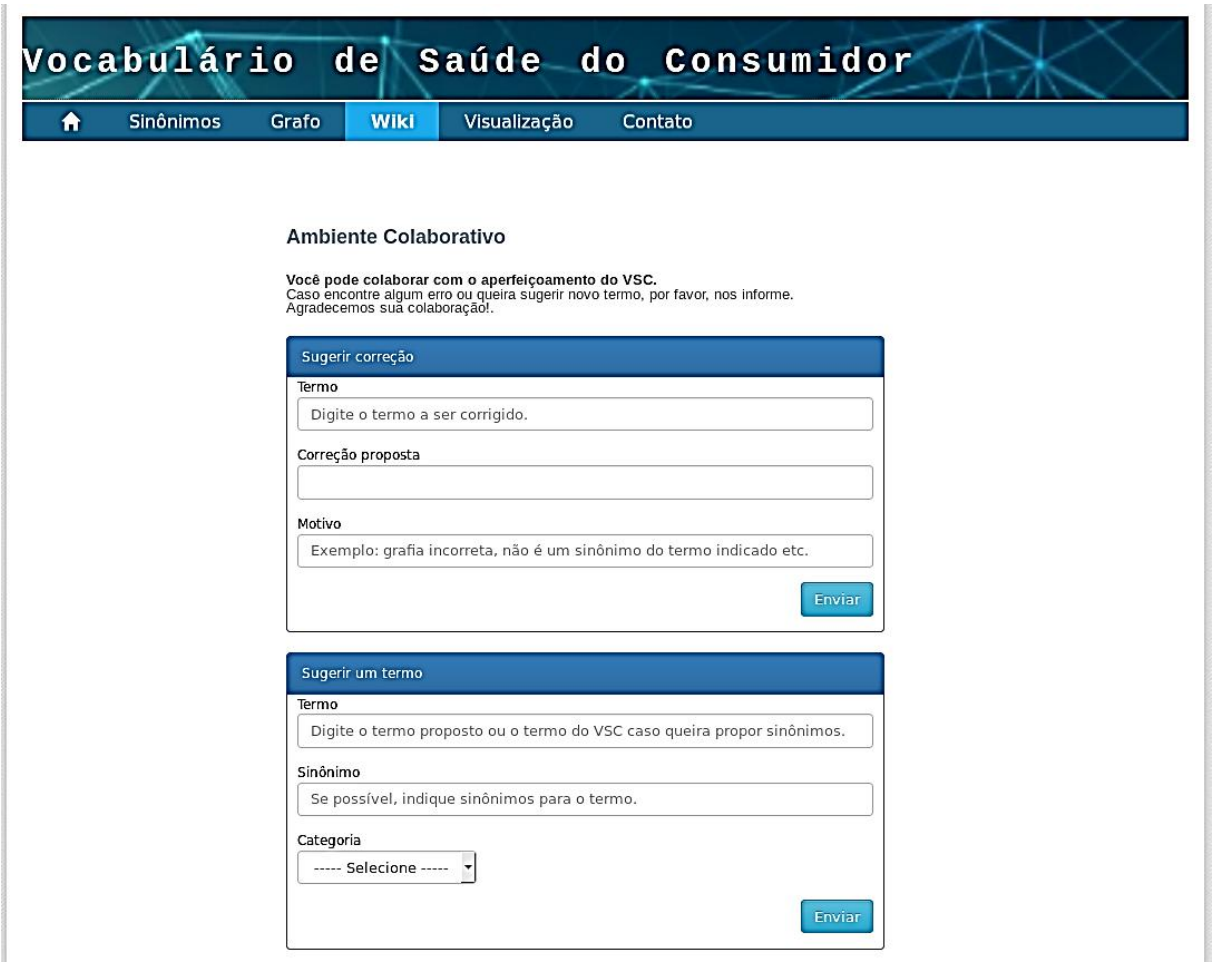


Figura 47. Interface colaborativa para atualização do VSC.

6 DISCUSSÃO

A primeira geração da OAC-CHV proposto por Zeng et al.[9], disponível em idioma inglês, poderia ser considerada suficiente para a construção de vocabulários em outros idiomas a partir da tradução dos termos, visto que os termos disponíveis pelos vocabulários da UMLS são identificados pelo CUI.

Todavia, o estudo de Qenam[103], ao analisar um conjunto de termos relacionados à radiologia, mostrou que a OAC-CHV tem ampla cobertura dos termos de busca, porém vários termos de uso comum não foram encontrados, o que pode impactar seu uso em aplicações disponíveis para consumidores em saúde. No estudo de Zeng-Treitler et al.[113] uma máquina de tradução automática foi utilizada para realizar a tradução de registros médicos. Verificou-se que a simples tradução não foi suficiente para conferir significado ao texto.

A proposta desse estudo foi utilizar os termos da OAC-CHV como sementes para busca em outras bases disponíveis no idioma português. Esta estratégia foi utilizada pelo fato de que os termos da OAC-CHV constituírem um subconjunto pré-classificado como pertencente ao domínio do consumidor. Esse método pode ser aplicado a bases disponíveis em outros idiomas, porém uma análise prévia da consistência dos dados deve ser realizada.

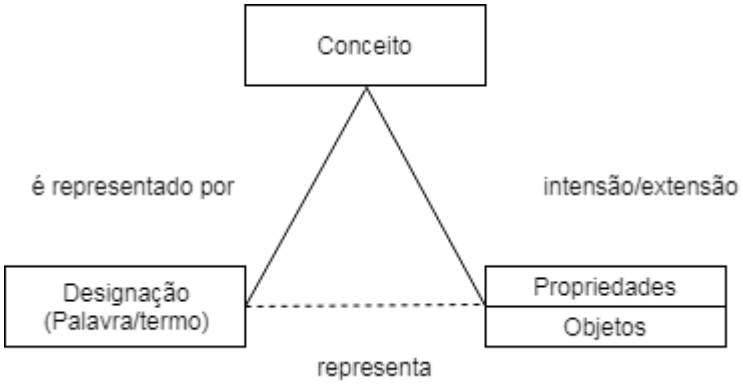
Outra possibilidade para a condução desse estudo seria apenas realizar a conexão de bases de dados. Os dados resultantes da intersecção UMLS/OAC-CHV/DeCS, suportados por uma ontologia e disponíveis em um modelo de dados adequado, seriam conectados com os dados da DBpedia. As consultas aos recursos poderiam ser realizadas por meio de um *endpoint* SPARQL[114], que propiciariam recuperar sinônimos, termos relacionados (*links*) e outras propriedades. Essa solução é a mais indicada para acesso aos dados, porém ao recuperar propriedades como termos relacionados (*links*), além da grande quantidade, não há nenhuma avaliação sobre a importância desses e nenhum apoio para a continuidade da exploração de dados realizada pelo usuário.

Em relação ao conteúdo e estrutura do VSC, obteve-se um vocabulário com 150.995 termos, um valor substancial e próximo à quantidade de termos disponibilizados na primeira versão da OAC-CHV. O VSC suportado por um modelo

de rede complexa e no modelo RDF possibilitou mapear relações semânticas mais significativas para o uso dos dados que a relação de sinonímia da OAC-CHV e de outros modelos de CHV analisados. As relações de hiperonímia, por exemplo, podem ser mais significativas para os usuários que o próprio sinônimo do termo; podem associar um significado conhecido ao termo de consulta. Contudo, cerca de 30.000 termos devem ainda ser analisados porque não foram mapeados para as bases consultadas. Outro item importante é que ao consumidor em saúde foi garantida a informação sobre a origem do dado no modelo de VSC construído.

De acordo com o mapeamento realizado neste estudo e com a formalização da ontologia SKOS, o VSC poderia ser melhor definido como um vocabulário que mapeia relacionamentos entre termos, por meio de sinonímia, e entre os conceitos, representados pelos termos preferidos, por meio de hiperonímia e associação. Trata-se de uma visão mais ampla e adequada que a definição original dos CHVs.[11]

Neste estudo considerou-se a ideia de que conceitos substituem o pensamento, de acordo com o triângulo semiótico, e são representados por meio de termos. Conceitos são definidos por meio de sua extensão (denotação) e intensão (conotação). Segundo Stock [115], no contexto da ciência da informação o conceito é definido como uma classe que contém objetos que possuem determinadas propriedades, representadas por suas relações semânticas neste estudo (Figura 48).



Extraído e traduzido de Stock.[115]

Figura 48. Triângulo semiótico na visão da ciência da informação.

Stock[115] reitera a importância das relações semânticas, em especial paradigmáticas, como parte de interesse na representação de sistemas de conhecimento, o que geralmente não é considerado na construção desses

esquemas inclusive por normas técnicas, assim como a teoria do conceito. As relações paradigmáticas são classificadas como de equivalência (como os sinônimos), hierarquia (hipônimos e hiperônimos) e por associação (termos relacionados). Outro aspecto importante apontado é que, dentre os sistemas de organização de conhecimento, como tesauros, as ontologias são estruturadas para representar de forma completa as relações semânticas. Neste estudo a representação das relações semânticas do VSC foi suportada por uma ontologia específica e pelo modelo de dados RDF.

6.1 Modelo de rede complexa de termos controlados-consumidor

A utilização de redes complexas possibilitou a análise dos termos recuperados e da associação entre estes, indicando a validade da estrutura proposta para suportar o desenvolvimento de aplicações para consumidores em saúde, como proposto por Zeng.[9] Gefen et al.[116] mostraram a importância dos estudos sobre a associação entre termos de saúde e indicaram que podem ser úteis para revelar casos que exigem especial atenção.

A combinação entre os vocabulários do UMLS e da base de conhecimento DBpedia foi proposta em outros estudos, que confirmam a importância da DBpedia como um meio acessível e importante para a recuperação de termos do consumidor. Mrabet et al.[28] mostraram uma melhoria significativa mensurável na recuperação de termos extraídos e identificação de sua categoria ao combinar estas bases. O presente estudo reforça o pressuposto presente na conclusão de Mrabet, que indica que a categorização é importante para a compreensão de conteúdos gerados pelo usuário. Os conteúdos foram obtidos a partir da combinação destas bases, que resultou no aumento da interação entre termos e consequente melhoria da identificação das categorias dos termos, por meio do mapeamento direto ou por meio da estrutura hierárquica da rede complexa. Em outra abordagem, Tilahun et al.[29] utilizaram tecnologias da web semântica para conectar bases de conhecimento de saúde disponíveis na web, como a DBpedia, e criar uma visualização de dados como um meio para o enriquecimento de dados.

A categorização dos termos e a estrutura de rede que suporta a construção

da Fase 1 do VSC também são importantes para lidar com situações de polissemia e homonímia, visto que o relacionamento do termo com seu hiperônimo ou hipônimo pode revelar seu significado.

A despeito do fato de a DBpedia ser construída por meio de extratores que convertem partes específicas do conteúdo da Wikipédia escrito por e para consumidores[68], este estudo mostrou que cerca de seis mil termos exatos do UMLS (9,1%) estão presentes nesta, sendo que 36% destes são termos técnicos. Apesar do baixo percentual, esses valores foram considerados relevantes porque esses termos possibilitaram a interligação entre as redes, além de propiciar extração de cerca de 40.000 termos anotados por similaridade semântica ou parte do termo, o que ampliou o relacionamento entre os termos de buscas e os anotados. Esta integração também resultou em um aumento expressivo do número de sinônimos e outras estruturas semânticas relacionadas a estes.

A identificação de sinônimos é uma difícil tarefa computacional [117] pelo fato da inexistência de uma relação clara entre extração de termos e suas relações.[118] Tapi-Nzali et al.[22] apresentaram um método complexo de extração de sinônimos para construção de um CHV, porém este modelo foi restrito a recuperação de termos referentes apenas a câncer de mama. O processo de recuperação de sinônimos reportado no presente estudo indicou que a DBpedia pode ser uma fonte importante e acessível de sinônimos previamente mapeados porque possibilitou ampliar o número médio de sinônimos referentes a cada termo preferido. Este estudo conseguiu identificar, por meio da DBpedia, aproximadamente, 24.000 pares termos preferidos-sinônimos dos vocabulário do UMLS. Vydiswaran [23] utilizou a Wikipédia para identificar sinônimos utilizando termos de ligação entre frases e encontrou 2.721 pares.

A utilização da DBpedia como fonte de sinônimos e acrônimos mostrou-se importante para obtenção desses tipos e contribuiu para o aumento da média de sinônimos por termo preferido e identificação de sinônimos mais populares. Neste estudo a propriedade *wikiPageRedirects*, pertencente à ontologia da DBpedia, foi utilizada como sinônimo, apesar de serem considerados como *redirects* termos com significado apenas próximo ao termo. A utilização de *redirects* como sinônimo foi considerada apropriada porque na ausência de um sinônimo estrito, um termo com

significado próximo pode ser importante para a compreensão do conceito, assim como para fins de navegação se usado em aplicações.[70]

Outro aspecto importante do uso da DBpedia foi a ampliação do número médio de associações entre termos, inclusive entre termos de outras categorias. Um dos efeitos da estrutura proposta foi o mapeamento inédito, de acordo com as bases descritas, entre conceitos estabelecidos por consumidores em saúde. Por exemplo, os termos “hemácias” (organismo) e “isquemia” (doenças) são comuns a ambas as redes e estão relacionados mesmo pertencendo a categorias diferentes. A estrutura proposta possibilitou mapear essa relação e visualizá-la por meio da análise de agrupamentos, apresentados na Tabela 2 (p. 48). Essa associação é comumente descrita em livros de medicina [119], mas não por consumidores em saúde. Outro exemplo são as interações mapeadas para o termo “gripe”, relacionando-o, inclusive a “antibióticos” e a eventos históricos.

Apesar da variabilidade de assuntos que emergiram da base DBpedia, a Tabela 2 mostrou a predominância dos termos associados a “doenças”, “organismos” e “localização geográfica”, um resultado coerente com uma base de domínio da saúde. A Tabela 3 (p. 49) corrobora a predominância de termos de saúde nas redes complexas construídas, indicada pelas métricas associadas à importância dos termos.

A DBpedia é uma base escrita por e para consumidores, porém a grande quantidade de termos e a diversidade dos assuntos encontrados, muitas vezes não relacionados à saúde, implicam que a possibilidade de recuperar simplesmente seu conteúdo, a partir da tecnologias da web semântica, possa ser potencialmente problemática caso não haja uma análise prévia. A proposta apresentada nesta tese é que um método automático seja aplicado para eleger termos candidatos ao vocabulário, o que foi realizado por meio de um processo semiautomatizado, baseado em uma abordagem não supervisionada e refinado por meio de seleção manual.

Verificou-se que a aplicação do algoritmo Louvain adotado foi suficiente para conferir a separabilidade do conjunto de termos candidatos ao VSC. Todos os termos coletados ou extraídos dos vocabulários OAC-CHV/MeSH/DeCS foram mantidos e foi possível separar termos característicos da área de saúde

provenientes da DBpedia daqueles referentes a assuntos diversos.

Um dos fenômenos observados empiricamente foi a revelação de agrupamentos gerados por meio do algoritmo Louvain, que formaram estruturas caracterizadas por assuntos, analogamente a um modelo de tópicos.[120, 121] Este relacionamento entre termos também pode ser um item a ser utilizado no desenvolvimento de novas aplicações, como uma ferramenta de auxílio à escrita e/ou verificação sobre um tema específico que ofereça um conjunto de termos característico.

Para os resultados aqui apresentados, em que há intenção de se criar uma infraestrutura de dados para suportar futuras aplicações, o modelo de rede complexa para suportar o VSC mostrou-se válido por atribuir importância e utilização prática ao conjunto de termos, à estrutura semântica (Figuras 15 a 17, p. 53 a 55) e às mensurações características. Todos esses aspectos podem ser a base para a realização de pesquisas e a construção de artefatos que suportam as necessidades dos consumidores em saúde.

6.2 Análise de conteúdos não estruturados

O processo de busca por termo exato em conteúdos não estruturados, essencialmente jornalísticos, escritos por ou para pacientes ou consumidores em saúde, mostrou-se eficiente para identificar termos pertencentes aos vocabulários do UMLS em conjunto com a medida da frequência destes, que foi utilizada neste estudo para fins de organização e apresentação destes (Tabelas 4 a 6, p. 59 e 60). Os termos do UMLS que ocorreram nos conteúdos pesquisados foram considerados válidos e compuseram o VSC. Considerou-se o pressuposto que, se o termo foi utilizado mesmo raramente, para comunicação com o consumidor, então o VSC deve oferecer suporte semântico. Em relação à medida de frequência, esta pode ser importante na continuidade do estudo, para apoiar análises sobre popularidade ou familiaridade de termos de saúde.

Em relação aos conteúdos analisados, não foram encontrados estudos que relatassem a utilização de notícias, nem a quantidade analisada no estudo aqui apresentado. Os estudos disponíveis na literatura pesquisada descreveram a busca

por termos em conteúdos como redes sociais [16, 22, 34, 96, 97] e registro eletrônico do paciente.[33, 99]

A busca por termos do UMLS resultou em uma quantidade equiparável ou ligeiramente superior (7.550 termos - 5.916 distintos) ao estudo de Park [96] que analisou conteúdos referentes à categoria “diabetes”, no qual foram identificados 1.130 (MeSH) e 1.931 (OAC-CHV) termos em blogs, 2.957 (MeSH) e 4.928 (OAC-CHV) termos extraídos de perguntas publicadas em rede social e 4.796 (MeSH) e 7.625 (OAC-CHV) termos em respostas associadas. Não foi reportada a ocorrência de sobreposições de termos, o que pode indicar que a quantidade de termos distintos possa ser menor que os valores indicados. Ressalta-se que a versão original dos vocabulários MeSH e OAC-CHV contém mais termos que a versão em português usada neste estudo. Os termos encontrados por este estudo neste processo pertencem às bases MeSH e OAC-CHV. As categorias mais frequentes foram, em tradução livre, “Aminoácidos, peptídeos ou proteínas” (Compostos químicos e drogas), “Partes do corpo ou órgão” (Anatomia) e “Doenças ou síndromes" (Doenças). Esses resultados são análogos aos apresentados na Tabela 5 (pág. 60).

No presente estudo a quantidade de temos do UMLS e OAC-CHV foi inferior ao de He et al.[34] De fato, He utilizou um método computacional mais elaborado que Park e identificou cerca de 7.000 termos distintos pertencentes a OAC-CHV, sendo que 2.465 termos do UMLS, mas não mapeados na OAC-CHV, em conteúdos referentes a perguntas e respostas sobre “diabetes”. Valores análogos, sendo cerca de 8.800 termos OAC-CHV distintos e 2.895 termos do UMLS, mas não OAC-CHV foram encontrados a partir de perguntas e respostas sobre “câncer”.

A identificação de aproximadamente 1.000 termos dos vocabulários do UMLS que não pertenciam a OAC-CHV foi importante para aumentar a cobertura da OAC-CHV em português. Estes são termos importantes por estarem diretamente relacionados ao uso para comunicação com o consumidor.

A aplicação de técnicas de análise de texto para fins de identificação de termos importantes, nesse caso para compor vocabulários de saúde do consumidor, foi descrita em diferentes estudos. Os algoritmos descritos são baseados em análise linguística e estatística [16, 22, 30] e aprendizagem não supervisionada para fins de

classificação.[33, 99] O algoritmo RAKE, apesar de ser construído para identificar palavras-chave no texto, mostrou-se efetivo porque apresenta uma estrutura análoga aos algoritmos ou métodos utilizados nesses estudos.

As medidas de efetividade do algoritmo RAKE, como precisão, *F-measure* e área sob a curva ROC (AUC-ROC), foram obtidas após a validação humana, o que possibilitou uma comparação com os estudos análogos.

No presente estudo obteve-se 0,57 (moderado) para AUC-ROC, um valor inferior aos reportados em outros estudos [30, 33, 99] que reportaram valores entre [0,70 - 0,95] (bom – excelente). Todavia, ressalta-se que, em estudos de recuperação da informação, a curva de precisão-revocação é mais robusta, assim como medidas como *F-measure*, e precisão. Os resultados mostraram que a probabilidade mínima de o algoritmo RAKE identificar um termo de saúde válido foi 0,732**.** Este valor alcança cerca de 0,900 para cerca de 10% dos termos com maior índice (Figura 39, p. 78), o que indica que RAKE pode ser usado de forma eficaz para identificar termos do domínio da saúde, de forma análoga a outras técnicas mais refinadas ou que exijam processos mais elaborados. Outro fator importante é a velocidade de execução de RAKE e a facilidade de uso na execução da tarefa de identificação de termos.

A medida de precisão obtida [0,662 – 0,781] com valor médio 0,726 (Tabela 9, p. 77) foi superior ao valor obtido com técnicas como tf-idf e c-value, apresentados nos experimentos conduzidos por He et al.[34], e análogo ao valor obtido pelo algoritmo de teste simiTerm, que atingiu valores entre [0,70 – 080]. Chen et al.[101] alcançaram maior precisão (0,850) extraindo termos válidos de prontuários eletrônicos.

Considerando o valor de *F-measure* obteve-se no presente estudo valores entre [0,00 - 0,797] (Tabela 9, p. 77), com valor médio 0,547. Estes valores foram superiores aos alcançados nos experimentos conduzidos por He et al.[34], observados na distribuição do gráfico apresentado.

A identificação de sinônimos é uma difícil tarefa computacional pelo fato da inexistência de uma relação clara entre extração de termos e relações.[117, 118] Uma possibilidade é a utilização de padrões léxico-sintáticos, por exemplo, termo ligados em um texto por "também conhecidos como".[23] O estudo para a construção do CHV

francês descreve a identificação de sinônimos como abreviaturas e variações de escrita, utilizando diversos algoritmos baseados na avaliação de similaridade.[22]

No presente estudo a proposta para a identificação de sinônimos inéditos em referência aos vocabulários do UMLS foi baseada em duas estratégias: na coleta diretamente por meio da DBpedia e na avaliação humana do relacionamento entre os n-gramas e o termo preferido, associados por coocorrência. A análise realizada nas Figuras 33 a 35 (p. 71 e 72) indicou que RAKE pode apresentar n-gramas compostos por termos preferidos e sinônimos. Neste estudo, foram identificados manualmente termos com esta característica, o que indica que sinônimos inéditos podem ser obtidos por meio desta análise, apesar de não se tratar de um processo completamente automático e exigir algum esforço humano.

Outro item importante desse estudo foi a identificação, por meio do algoritmo RAKE, de termos exclusivos utilizados no sistema de saúde brasileiro, como “sistema único", “UPA” (Unidade de Pronto Atendimento) e “cenário assistencial-hospitalar”. Outros exemplos constam de termos popularizados recentemente, como “febre chikungunya”, “medicamento similar”, “distrito sanitário”, “compulsão alimentar”, que ainda não estão mapeados nos vocabulários do UMLS.

Apesar dos esforços na atualização destes vocabulários, termos técnicos usados na comunicação com o consumidor que tratam de assuntos de saúde emergentes ou específicos como “desordem genética autossômica recessiva”, “doença mitocondrial grave herdada”, “hiperidrose emocional”, “ressonância magnética cerebral”, “genoma embrionário humano”, “vertigem posicional paroxística benigna”, “complemento alimentar”, “alteração genética”, não foram detectados nos vocabulários controlados analisados, porém foram identificados e validados nesta fase. RAKE. Portanto, a estratégia foi eficiente para identificar termos que foram validados com até 4-gramas, o que não foi apresentado em estudos análogos.

6.3 Ontologia SKOS e modelo RDF

As ontologias SKOS e PAV foram satisfatórias para formalizar as relações semânticas entre os termos do VSC e os metadados associados ao vocabulário e termo de forma mais abrangente que a ontologia específica para CHV publicada por

Amith e Tao.[24] Apesar de a ontolologia MuEVo [89] apresentar maior completude na formalização do processo de aquisição de sinônimos, esta ontologia não foi usada porque este item não foi contemplado nesse estudo.

Uma característica importante à formalização do VSC é o mapeamento de relações hierárquicas entre termos, a exemplo de "Febre" → "Condições Patológicas, Sinais e Sintomas" → "Doenças". SKOS suporta inferência entre os termos ligados pelo predicado *skos:broaderTransitive.* A transitividade de skos: *broaderTransitive* faz com que a declaração "Febre" *skos:broaderTransitive* "Doenças" seja inferida. Esta característica foi importante por associar a categoria a termos não categorizados na origem, a exemplo de sinônimos coletados da DBpedia que somente estavam associados a um termo preferido. Stock [115] reitera a importância da transitividade para sistemas de organização do conhecimento e que esta característica não está contemplada em normas técnicas para construção destes.

Destaca-se que Bushman et al.[36] descreveram os esforços para a criação de uma representação dos dados do MeSH-RDF (id.nlm.nih.gov/mesh/). A organização dos dados nesse modelo possibilitou que o relacionamento entre os termos preferidos fosse mais bem estruturado de forma que as relações intrínsecas presentes na árvore hierárquica fossem representadas. O uso das tecnologias da web semântica possibilitou resultados promissores na integração de recursos heterogêneos. No presente estudo, essas vantagens também estão associadas ao VSC-RDF.

As tecnologias da web semântica são úteis nos cuidados à saúde, especialmente considerando que o tratamento de um paciente pode envolver várias referências a conceitos de terminologias como a CID-10 ou SNOMED. A construção do VSC-RDF foi pautada pela possibilidade de conectá-lo a outras bases da área de saúde publicadas na web [122] para fins de enriquecimento de dados, o que pode ampliar as potencialidades do VSC para fins de suporte ao desenvolvimento de aplicações que auxiliem na resolução de demandas dos consumidores em saúde.

6.4 Atualização do VSC

A aquisição de temos, que propicia ampliar a cobertura dos vocabulários, é uma necessidade constante. Métodos automáticos são usados na tarefa de identificar termos, porém, identificar termos do domínio da saúde requer que questões como a confiabilidade sejam mais bem observadas de forma que algum esforço humano para validar os termos seja necessário.

Neste estudo o algoritmo RAKE identificou aproximadamente 91.000 termos candidatos, mas devido à necessidade de validação manual, apenas cerca de 10% desses termos foram analisados. Há, portanto, cerca de 80.000 termos candidatos disponíveis para avaliação na continuidade do estudo. Métodos de validação automática poderão ser aplicados para minimizar o impacto da necessidade de esforço humano.

Outra possibilidade é a aplicação do mesmo processo executado na Fase 2 na exploração de outros conteúdos que versam sobre saúde, a saber, *tweets*, redes sociais com foco na saúde, ou mesmo o conteúdo e relações mapeados pelo Google Trends (trends.google.com.br/trends/).

Além disso, as contribuições continuamente atualizadas da Wikipédia são úteis para a ampliação do VSC especificamente para novos termos pertencentes ao domínio da saúde. Ngo et al.[123] realizaram um estudo sobre a completude e adequação da Wikipedia na criação de uma terminologia clínica e concluíram que 80% dos conceitos do SNOMED estão cobertos pelos títulos disponíveis.

As contribuições dos consumidores em saúde também compõem as estratégias de atualização e confiabilidade do conteúdo do VSC, baseadas na disponibilização do ambiente colaborativo (Figura 47, p. 88).

6.5 Limitações

Apesar da quantidade de termos recuperada do MeSH, apenas foram recuperados os termos referentes a algumas estruturas semânticas. Estruturas disponibilizadas na UMLS MethamorphoSys como *ancestors, siblings, descendants and pharmacological action list* não foram mapeadas devido à dificuldade computacional

de manipulação da grande quantidade de termos.

Em relação à relevância da DBpedia como fonte de dados, este estudo mostrou que uma análise prévia dos dados deve ser realizada de forma a excluir relacionamentos semânticos múltiplos, acrônimos não condizentes com o significado no domínio da saúde e unificar os termos com variações de acentuação. Naturalmente para um conjunto maior de dados devem ser utilizados meios automáticos para esse processo.

Outro item importante associado à DBpedia é o fato de que em sua ontologia a propriedade *wikiPageRedirects* é referente a sinônimos e quase-sinônimos. Liu[70] considera que 78% desses termos são efetivamente sinônimos, porém para este a presença de quase-sinônimos associados por equivalência a um termo preferido por causa da herança dos não representa uma incorreção, visto que os quase-sinônimos podem ser importantes para a navegação entre termos.

Um problema ocorreu ao recuperar termos cuja estrutura semântica não foi mapeada. A organização dos termos em rede e o VSC-RDF possibilitaram o estabelecimento dessa estrutura para um conjunto de termos, porém, ainda há termos cuja estrutura não está completa, o que pode influenciar na efetividade de aplicações baseadas no modelo de VSC.

Em relação à análise dos conteúdos não estruturados, o método aplicado limitou-se a identificação de termos exatos, o que diminuiu a contagem de termos do UMLS encontrados.

Devido à grande quantidade de n-gramas identificados pelo algoritmo RAKE foi necessário atribuir um valor de corte arbitrário para considerar os n-gramas candidatos. Naturalmente, dentre os n-gramas não considerados, é possível que tenham sido excluídos do processo termos de saúde potencialmente importantes para a primeira versão do VSC.

Para a execução do algoritmo RAKE foi necessária uma análise linguística (etiquetagem POS) prévia para filtrar categorias gramaticais. A estrutura composta por substantivo e adjetivo apresentou n-gramas melhor identificados com termos de saúde. A estrutura dos n-gramas obtidos é restrita a estas categorias gramaticais. Desta forma, estruturas importantes compostas por também por conjunções não foram identificadas, porém são presentes e importantes para o idioma português.

6.6 Trabalhos futuros

O objetivo subjacente desse estudo foi a criação de uma infraestrutura de dados relacionáveis segundo uma estrutura semântica básica e formalizada de acordo com princípios de compartilhamento e disponibilidade, itens preconizados pela web semântica.

A continuidade do estudo prevê o armazenamento dessa infraestrutura em um servidor de triplas, de forma a propiciar o uso ampliado das tecnologias da web semântica, como o enriquecimento de dados baseado na conexão de bases abertas (*linked open data*).[122] Ao estabelecer uma conexão com outras bases diferentes aplicações podem ser exploradas, como a conexão de termos do consumidor a dados de genômica, geográficos ou de linguística, o que poderia ampliar o número de sinônimos, potencializar visualizações e a aplicação de métodos para a identificação de fenômenos.

Diante desse cenário outras possibilidades para este estudo contam com a condução de experimentos para fins de mensuração de popularidade ou familiaridade dos termos, além da avaliação de itens como a confiabilidade dos termos extraídos de vocabulários não controlados.

7 CONCLUSÃO

Este estudo mostrou que os conteúdos estruturados e não estruturados disponíveis na web compõem uma base útil para a construção de um modelo de vocabulário de saúde com foco no consumidor.

Segundo Zeng [30] há dois princípios que guiam o esforço de identificação de termos para construção de CHV: devem ser compostos por termos reais comumente usados por consumidores e devem possibilitar o processamento da linguagem por computador.

O princípio referente a termos reais dos consumidores foi contemplado pela utilização da base de conhecimento DBpedia e pelos conteúdos não estruturados escritos por e para consumidores na Wikipedia.

A DBpedia mostrou-se uma importante fonte de sinônimos e outros relacionamentos entre termos, apesar da necessidade de uma análise prévia do conjunto e das relações entre termo, o que foi feito por meio do modelo de rede complexa. Trata-se de uma base de fácil acesso, atualizada com frequência e criada colaborativamente por usuários ou consumidores em saúde. Outro aspecto importante é possibilidade de conexão desses termos com outras bases de dados, promovendo o enriquecimento da estrutura do vocabulário.

A estruturação dos termos em redes complexas mostrou-se importante para verificar a consistência do conjunto de dados por meio da análise das medidas específicas. A utilização da rede complexa para representação das relações entre termos possibilitou estas sejam analisadas do ponto de vista dos vocabulários controlados e do consumidor em saúde, de forma a apresentar características referentes às associações entre termos ainda não estabelecidos diretamente. Além disso, a estrutura possibilita a navegação entre termos, base para o desenvolvimento de aplicações com essa característica.

A utilização de conteúdos não estruturados disponíveis na web, em conjunto com a aplicação de técnicas de análise de texto, pode colaborar com a extração de termos inéditos para a construção de um CHV em idioma português. Essa estratégia mostrou-se válida para a identificação de termos de domínio do consumidor, além da possibilidade de mapear termos exclusivos do sistema de saúde local.

O segundo princípio, referente ao processamento por computador, foi contemplado pela proposta de disponibilizar os termos, relacionamentos e metadados por meio de um modelo de dados, o RDF, que possibilita a conexão entre dados.

Apesar das limitações, o modelo de VSC proposto mostrou-se válido como uma estrutura baseada no conceito de vocabulário da web semântica ao possibilitar a recuperação de sinônimos mais populares e mostrar seu potencial para a análise de fenômenos de interação entre termos estabelecidos por consumidores em saúde.

Os resultados obtidos mostraram que a web pode ser uma fonte para a prospecção de termos e relações, visto que a partir da Web 2.0 os usuários passaram a produzir conteúdo. O rastreamento desse conteúdo, neste estudo, nos levou a mapear os diversos usos de sinônimos e relacionamentos entre termos feitos pelo consumidor em saúde.

As limitações da aplicação dessa técnica para a obtenção de conteúdo se encontram no aspecto da confiabilidade dos dados recuperados de fontes do consumidor. Por isso, a validação humana é ainda um passo importante para a qualidade do resultado.

A essência desse estudo é a criação de uma infraestrutura de dados disponíveis e conectáveis, compondo uma base para o desenvolvimento de aplicações.

O esforço e escolhas realizadas no desenvolvimento do VSC foram orientados pelas potencialidades dessas aplicações. Todavia, desenvolver aplicações funcionais não faz parte do escopo desta tese.

As aplicações em desenvolvimento constam do suporte à leitura e compreensão de relatórios médicos e aos registros eletrônicos em saúde. Outras possibilidades são:

- identificar termos técnicos, oferecer sinônimos e termos relacionados em prescrições médicas ou bulas lidas por meio de dispositivos móveis;
- auxiliar a escrita e/ou verificação de temas específicos, indicando um conjunto de termos característicos do tema;
- desenvolver uma ferramenta para reconhecimento de conceitos, análogo ao Metamap (metamap.nlm.nih.gov);

- disponibilizar serviços para suportar pesquisas e exploração dos conceitos e associações estabelecidas pelo consumidor;
- criar serviços que suportem a conexão dos dados do VSC a outras bases, por exemplo, de dados genéticos, instituições, especialistas para que a navegação possa ampliar o escopo inicial da busca.

Esperamos que esse estudo contribua como um método para a construção de vocabulários de saúde do consumidor em outros idiomas, além de apresentar-se como uma possibilidade de atualização dos vocabulários existentes.

8 REFERÊNCIAS

1. Zeng Q, Kogan S, Ash N, Greenes RA. Patient and clinician vocabulary: how different are they? Stud Health Technol Inform. 2001;84(Pt 1):399–403.

2. Keselman A, Logan R, Smith CA, Leroy G, Zeng-Treitler Q. Developing informatics tools and strategies for consumer-centered health communication. J Am Med Inform Assoc. 2008;15(4):473–83.

3. Pew Internet and American Life Project. The online health care revolution: how the web helps Americans take better care of themselves [Internet]. 2000 [cited 2018 Mar 4]. Available from: http://www.pewinternet.org.

4. Eysenbach G, Köhler C.How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews.BMJ. 2002 Mar 9;324(7337):573-7. DOI: https://doi.org/10.1136/bmj.324.7337.573.

5. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. Med J Aust. 2018 Aug 20.

6. Chen YY, Li CM, Liang JC, Tsai CC. Health information obtained from the internet and changes in medical decision making: questionnaire development and cross-sectional survey. J Med Internet Res. 2018;20(2):e47. DOI: 10.2196/jmir.9370.

7. Tanabe K, Fujiwara K, Ogura H, Yasuda H, Goto N, Ohtsu F Quality of web information about palliative care on websites from the United States and Japan: comparative evaluation study. Interact J Med Res. 2018 Apr 3;7(1):e7. DOI: 10.2196/ijmr.9574.

8. Boyer C, Gaudinat A, Hanbury A, Appel RD, Ball MJ, Carpentier M, et al. Accessing reliable health information on the web: a review of the HON approach. Stud Health Technol Inform. 2017;245:1004-1008.

9. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc. 2006 Jan 1;13(1):24–9.

10. Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC. Exploring lexical forms: first-generation Consumer Health Vocabularies. AMIA Annu Symp Proc. 2006; 2006:1155.

11. Open Access, Collaborative Consumer Health Vocabulary Initiative [Internet]. [cited 2018 Mar 9]. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html.

12. 2011AA Consumer Health Vocabulary Source Information [Internet]. 2010 [cited 2018 Mar 9]. Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/2011AA/CHV/mrsab.html.

13. Zeng Q, Kogan S, Ash N, Greenes RA, Boxwala AA. Characteristics of consumer terminology for health information retrieval. Methods Inf Med. 2002;41(4):289–98.

14. Zeng QT, Tse T, Crowell J, Divita G, Roth L, Browne AC. Identifying Consumer-Friendly Display (CFD) names for health concepts. AMIA Annu Symp Proc. 2005;2005:859–63.

15. Zeng-Treitler Q, Goryachev S, Tse T, Keselman A, Boxwala A. Estimating consumer familiarity with health terminology: a context-based approach. J Am Med Inform Assoc. 2008 Jun;15(3):349–56.

16. Doing-Harris KM, Zeng-Treitler Q. Computer-assisted update of a consumer health vocabulary through mining of social network data. J. Med. Internet Res. 2011;13(2):e37.

17. Sousa FS, Teixeira F, Nunes FLS, Domenico EBL. Abordagens para construção de vocabulários de saúde para o consumidor: uma revisão da literatura. In: XIII Congresso Brasileiro de Informática em Saúde - CBIS 2012, Anais do XIII Congresso Brasileiro de Informática em Saúde. Sociedade Brasileira de Informática em Saúde – SBIS, Curitiba 2012,PR.

18. Smith CA. Consumer language, patient language, and thesauri: a review of the literature. J Med Libr Assoc. 2011 Apr;99(2):135-44.

19. W3C 2015. Semantic web [Internet]. 2015 [cited 2018 May 2]. Available from: w3.org/standards/semanticweb/.

20. W3C 2014. Resource Description Framework (RDF) [Internet]. 2014 [cited 2016 Mar 9]. Available from: w3.org/RDF/.

21. W3C 2015. What is a vocabulary? [Internet]. 2015 [cited 2018 May 9]. Available from: https://www.w3.org/standards/semanticweb/ontology.

22. Tapi Nzali MD, Aze J, Bringay S, Lavergne C, Mollevi C, Optiz T. Reconciliation of patient/doctor vocabulary in a structured resource. Health Informatics J. 2018 Jan 1 DOI: 10.1177/1460458217751014.

23. Vydiswaran VG, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. AMIA Annu Symp Proc. 2014 Nov 14;2014:1150-9. eCollection 2014.

24. Amith T, Tao C. Ontology of Consumer Health Vocabulary. [Internet]. 2016 [cited 2018 Mar 9]. Available from: http://bioportal.bioontology.org/ontologies/OCHV.

25. Commission of the European Communities. Multilingual glossary of technical and popular medical terms in nine European languages. Heymans Institute of Pharmacology. 1995.

26. Seedor M, Peterson KJ, Nelsen LA, Cocos C, McCormick JB, Chute CG, Pathak J. Incorporating expert terminology and disease risk factors into consumer health vocabularies. Pac Symp Biocomput. 2013:421-32.

27. Hou L, Kang H, Liu Y, Li L, Li J. Mining and standardizing chinese consumer health terms. BMC Med Inform Decis Mak. 2018 Dec 7;18(Suppl 5):120. DOI: 10.1186/s12911-018-0695-6.

28. Mrabet Y, Kilicoglu H, Roberts K, Demner-Fushman D. Combining open-domain and biomedical knowledge for topic recognition in consumer health questions. AMIA Annu Symp Proc. 2017 Feb 10;2016:914-923.

29. Tilahun B, Kauppinen T, Keßler C, Fritz F. Design and development of a linked open data-based health information representation and visualization system: potentials and preliminary evaluation. JMIR Med Inform. 2014 Oct 25;2(2):e31. DOI: 10.2196/medinform.3531.

30. Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC, Goryachev S, Ngo L. Term identification methods for consumer health vocabulary development. J Med Internet Res. 2007 Feb 28;9(1):e4.

31. Jiang L, Yang CC, Li J. Discovering consumer health expressions from consumer-contributed content. In: Greenberg AM, Kennedy WG, Bos ND, editors. Social Computing, Behavioral-Cultural Modeling and Prediction SBP 2013. Berlin, Heidelberg: Springer; 2013. p. 164-5.

32. Kim J, Joo J, Shin Y. An exploratory study on the health information terms for the development of the consumer health vocabulary system. Stud Health Technol Inform. 2009;146:785.

33. Chen J, Zheng J, Yu H. Finding important terms for patients in their Electronic Health Records: a learning-torRank approach using expert annotations. JMIR Med Inform. 2016 Nov 30;4(4):e40.

34. He Z, Chen Z, Oh S, Hou J, Bian J. Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach. J Biomed Inform. 2017 May;69:75-85. DOI: 10.1016/j.jbi.2017.03.016.

35. Zeng-Treitler Q, Goryachev S, Kim H, Keselman A, Rosendale D. Making texts in electronic health records comprehensible to consumers: a prototype translator. AMIA Annu Symp Proc. 2007 Oct 11:846-50.

36. Bushman B, Anderson D, Fu G. Transforming the Medical Subject Headings into linked data: creating the authorized version of MeSH in RDF. J Libr Metadata. 2015;15(3-4):157-76.

37. National Labrary of Medicine (U.S.). Unified Medical Language System® (UMLS®). UMLS Glossary. [Internet]. [updated 2012 Aug 15; cited 2018 Jun 10]. Available from: https://www.nlm.nih.gov/research/umls/new_users/glossary.html.

38. de Keizer NF, Abu-Hanna A. Understanding terminological systems. II: Experience with conceptual and formal representation of structure. Methods Inf Med. 2000 Mar;39(1):22-9.

39. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32. DOI: 10.1093/nar/gkh061.

40. Campbell KE, Oliver DE, Spackman KA, Shortliffe EH. Representing thoughts, words, and things in the UMLS. J Am Med Inform Assoc. 1998 Sep-Oct;5(5):421-31.

41. Amaglobeli G. Semantic triangle and linguistic sign. Scientific Journal in Humanities. 2012;1(1):37-40.

42. Obrst L. Ontological Architectures.In: Poli R, Healy M, Kameas A, editors. Theory and applications of ontology: computer applications: Netherlands: Springer; 2010. p. 27-66.

43. U.S. National Library of Medicine. Relationships in Medical Subject Headings (MeSH). [Internet]. [cited 2019 Jul 04]. Available from: https://www.nlm.nih.gov/mesh/meshrels.html.

44. Diesner J. ConText: Software for the integrated analysis of text data and network data. In: Conference of International Communication Association (ICA). 2014. Seattle, WA.

45. Holanda AJ, Pisa IT, Kinouchi O, Ruiz EES. Thesaurus as a complex network. Physica A. 2004 Nov; 344(3-4):p. 530-6. DOI: 10.1016/j.physa.2004.06.025.

46. Vandenbussche P-Y, Charlet J. Méta-modèle général de description de ressources terminologiques et ontologiques. In Gandon FL, editors. Actes d'IC. France, Paris: PUG; 2009, p. 193–204.

47. Nooy W, Mrvar A, Batagelj V. Exploratory social network analysis with Pajek. New York: Cambridge University Press; 2011.

48. Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. BioData Min. 2011 Apr 28;4:10. DOI: 10.1186/1756-0381-4-10.

49. Simões-Pereira, JMS. Grafos e redes: teoria e algoritmos básicos. Rio de Janeiro: Interciência; 2013. 354p.

50. Citraro S, Rossetti G. A complex network approach to semantic spaces: how meaning organizes itself. Proceedings of the 27th Italian Symposium on Advanced Database Systems. 2019. Articoli 1–2.

51. Batagelj V, Mrvar A, Zaversnik M. Network analysis of dictionaries. Ljubljana: Preprint series; 2002. ISSN 1318-4865.

52. Fortunato S. Community detection in graphs. Physics Reports.2010 Feb;486(3-5):75-174.

53. Ognyanova K. Network Analysis and Visualization with R and igraph. [Internet]. 2016 [cited 2019 Jul 20]. Available from: https://kateto.net/netscix2016.html.

54. Liu K, Hogan WR, Crowley RS. Natural Language Processing methods and systems for biomedical ontology learning. J Biomed Inform. 2011 Feb;44(1):163-79. DOI: 10.1016/j.jbi.2010.07.006.

55. Krauthammer M, Nenadic G. Term identification in the biomedical literature. J Biomed Inform. 2004 Dec;37(6):512-26.

56. Frantzi KT, Ananiadou S. Automatic Term Recognition using contextual cues. Proceedings of 3rd DELOS Workshop; 1997.

57. Siddiqi S, Sharan A. Keyword and keyphrase extraction techniques: a literature review. Int J Comput Appl. 2015; 109(2), 18–23.

58. Knoth P, Schmidt M, Smrz P, Zdrahal Z. Towards a framework for comparing automatic term recognition methods. In: Conference Znalosti; 4-6 Feb 2009; Brno, Czech Republic.

59. Spasić I1, Greenwood M, Preece A, Francis N, Elwyn G. FlexiTerm: a flexible term recognition method. J Biomed Semantics. 2013 Oct 10;4(1):27. DOI: 10.1186/2041-1480-4-27.

60. Ramos JE. Using tf-idf to determine word relevance in document queries [Internet]. 2003 [cited 2018 Jul 05]. Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.1424&rep=rep1&type=pdf.

61. Frantzi K, Ananiadou S, Mima H. Automatic recognition of multi-word terms: the c-value/nc-value method. Int J Digit Libr. 2000; 3(2):115-130.

62. Barrón-Cedeño A, Sierra G, Drouin P, Ananiadou S. An improved automatic term recognition method for Spanish. In: Gelbukh A, editor. Computational Linguistics and Intelligent Text Processing. Berlin, Heidelberg: Springer; 2009. p. 125–36.

63. Conrado MS, Rossi RG, Pardo T, Rezende SO et al. Applying transductive learning for automatic term extraction: the case of the ecology domain. In Informatics and Applications (ICIA). Second International Conference on. IEEE; 2013. p. 264–69.

64. Foo J, Merkel M. Using machine learning to perform automatic term recognition. In: Proceedings of the LREC 2010 Workshop on methods for automatic acquisition of language resources and their evaluation methods. European Language Resources Association; 2010. p. 49-54.

65. Rose S, Engel D, Cramer N, Cowley W. Automatic keyword extraction from individual documents. In: Berry MW, Kogan J, editors. Text mining: theory and applications. Chichester (UK): John Wiley & Sons; 2010. 222 p.

66. Lakatos E. Fundamentos de metodologia científica. 8th ed. São Paulo: Grupo Gen - Atlas; 2017.

67. Biblioteca Virtual em Saúde. Descritores em Ciências da Saúde: DeCS. [Internet]. ed. 2017. São Paulo (SP): BIREME / OPAS / OMS. 2017 [updated 2017 May; cited 2015 Jul 28]. Available from: http://decs.bvsalud.org.

68. Morsey M, Lehmann J, Isele R, Jakob M, Jentzsch A, Kontokostas D et al. DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia. Semantic Web. 2015; 6(2):167-95. DOI: 10.3233/SW-140134.

69. Mendes PN, Jakob M, García-Silva A, Bizer C. DBpedia Spotlight: shedding light on the web of documents. In: Proceedings of the 7th International Conference on Semantic Systems (I-Semantics 2011). Graz, Austria, September 2011.

70. Liu O. Relation discovery on the DBpedia Semantic Web. Grenoble Institute of Technology - ENSIMAG. May 2009 [cited 2018 Mar 9]. Available from: https://ensiwiki.ensimag.fr/images/0/05/Dbpedia-relation-discovery-article.pdf.

71. Mrvar A, Batagelj V. Analysis and visualization of large networks with program package Pajek. Complex Adapt Syst Model. 2016; 4(6). DOI 10.1186/s40294-016-0017-8.

72. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, Costa LdF et al. Clustering algorithms: A comparative approach. PLoS One. 2019; 14(1): e0210236. https://doi.org/10.1371/journal.pone.0210236.

73. Emmons S, Kobourov S, Gallant M, Börner K. Analysis of network clustering lgorithms and cluster quality metrics at scale.PLoS One. 2016 Jul 8;11(7):e0159161. DOI: 10.1371/journal.pone.0159161.

74. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. communities in large networks. J. Stat. Mech. 2008; P10008.

75. Schaub MT, Delvenne JC, Rosvall M, Lambiotte R. The many facets of community detection in complex networks. Applied Network Science. 2017;2:4.

76. MongoDB Manual. Text Search. [Internet]. 2018 [cited 2019 Jun 25]. Available from: https://docs.mongodb.com/manual/text-search/.

77. Welbersa K, Van Atteveldtb W, Benoit K. Text analysis in R. Commun Methods Meas. 2017:11(4); 245–65. DOI:10.1080/19312458.2017.1387238.

78. Wijffels J, Straka M, Straková J. udpipe - R package for tokenization, tagging, lemmatization and dependency parsing based on UDPipe. [Internet]. GitHub. [cited 2018 Sep 24]. Available from: https://github.com/bnosac/udpipe.

79. Straka M, Hajic J, Strakova J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In: Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association, Portoroz, Slovenia. 2016.

80. Straka M, Strakova J. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Vancouver, Canada. 2017 August 3-4: 88-9.

81. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005 May;37(5):360-3.

82. Lo Martire R. Reliability coefficients. [Internet]. 2017 [cited 2019 Feb 05]. Available from: https://cran.r-project.org/web/packages/rel/index.html.

83. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015; 10(3): e0118432. DOI: 10.1371/journal.pone.0118432.

84. Davis J, Goadrich M. The relationship between precision-recall and ROC curves. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA; 2006. p.233-40.

85. Teixeira FO. Saúde 360: aplicação dos princípios da web semântica para interrelacionar informação no domínio da saúde. São Paulo. Tese [Doutorado] – Universidade Federal de São Paulo. Escola Paulista de Medicina. Programa de Pós-Graduação em Gestão e Informática em Saúde; 2016.

86. W3C Working Group Note. SKOS Simple Knowledge Organization System Primer. [Internet]. 2009. [cited 2018 Aug 25]. Available from: https://www.w3.org/TR/skos-primer/.

87. Baker T, Bechhofer S, Isaac A, Miles A, Schreiber G, Summers E. Key choices in the design of Simple Knowledge Organization System (SKOS). J Web Semantics. 2013 May; 20: p. 35-49.

88. Pastor-Sanchez JA, Martinez FJM, Rodriguez-Munoz JV. Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Information Research.2009; 14(4).

89. Eholié S, Tapi Nzali MD, Bringay S, Jonquet C. MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. In: Paschke A, Burger A, Splendiani A, Marshall M, Romano P, editors. SWAT4LS: Semantic Web Applications and Tools for Life Sciences, Amsterdam, Netherlands. 2016.

90. Ciccarese P, Soiland-Reyes S, Belhajjame K, Gray AJ, Goble C, Clark T. PAV ontology: provenance, authoring and versioning. J Biomed Semantics. 2013 Nov 22;4(1):37. DOI: 10.1186/2041-1480-4-37.

91. Boettiger C, Mecum B, Krystalli A, Senderov V. rdflib: tools to manipulate and query semantic data [Internet]. GitHub. 2019. [cited 2019 Apr]. Available from: https://github.com/ropensci/rdflib.

92. Boettiger C. rdflib: A high level wrapper around the redland package for common rdf applications (Version 0.1.0). Zenodo. 2018 [cited 2019 Apr 02]. Available from: https://doi.org/10.5281/zenodo.1098478.

93. Boettiger C. A tidyverse lover's intro to RDF. 2019. [cited 2019 Apr 02]. Available from: https://cran.r-project.org/web/packages/rdflib/vignettes/rdf_intro.html.

94. SPARQL Working Group. SPARQL Query Language for RDF. [Internet]. 2008 [cited 2019 Jul 01]. Available from: https://www.w3.org/TR/rdf-sparql-query/.

95. Soldaini L, Yates A, Yom-Tov E, Frieder Ophir, Goharian N. Enhancing web search in the medical domain via query clarification. Information Retrieval Journal. 2016: 19(1);149-73.

96. Park MS, He Z, Chen Z, Oh S, Bian J. Consumers' use of UMLS concepts on social media: diabetes-related textual data analysis in blog and social Q&A sites.JMIR Med Inform. 2016 Nov 24;4(4):e41.

97. Hou L, Kang H, Liu Y, Li L, Li J. Mining and standardizing chinese consumer health terms. BMC Med Inform Decis Mak. 2018 Dec 7;18(Suppl 5):120. doi: 10.1186/s12911-018-0695-6.

98. Chen C, Huang E, Yan H.Detecting the association of health problems in consumer-level medical text. Journal of Information Science. 2018:44(1);3-14.

99. Jiang K, Chen T, Calix RA, Bernard GR. Identifying consumer health terms of side effects in twitter posts. Stud Health Technol Inform. 2018;251:273-276.

100. Gu G, Zhang X, Zhu X, Jian Z, Chen K, Wen D, Gao L, Zhang S, Wang F, Ma H, Lei J. Development of a Consumer Health Vocabulary by mining health forum texts based on word embedding: semiautomatic approach. JMIR Med Inform. 2019 May 23;7(2):e12704. DOI: 10.2196/12704.

101. Chen J, Jagannatha AN, Fodeh SJ, Yu H. Ranking medical terms to support expansion of lay language resources for patient comprehension of Electronic Health Record notes: adapted distant supervision approach. JMIR Med Inform. 2017 Oct 31;5(4):e42. DOI: 10.2196/medinform.8531.

102. Chen J, Yu H. Unsupervised ensemble ranking of terms in electronic health record notes based on their importance to patients. Journal of Biom Inform. April 2017; 68:121-131.

103. Qenam B, Kim TY, Carroll MJ, Hogarth M. Text simplification using Consumer Health Vocabulary to generate patient-centered radiology reporting: translation and evaluation. J Med Internet Res. 2017 Dec 18;19(12):e417. DOIi: 10.2196/jmir.8536.

104. Shivade C, Malewadkar P, Fosler-Lussier E, Lai AM. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes.J Biomed Inform. 2015 Dec;58 Suppl:S103-10. DOI: 10.1016/j.jbi.2015.08.025.

105. Lopes CT, Ribeiro C. Identification and classification of health queries: co-occurrences vs. domain-specific terminologies (Report). Int J Healthc Inf Syst Inform. 2014;9(3);55(17).

106. Chen J, Druhl E, Polepalli Ramesh B, Houston TK, Brandt CA, Zulman DM, Vimalananda VG, Malkani S, Yu H. A Natural Language Processing system that links medical terms in Electronic Health Record notes to lay definitions: system development using physician reviews. J Med Internet Res. 2018 Jan 22;20(1):e26. DOI: 10.2196/jmir.8669.

107. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med. 2005 May;37(5): 360-3.

108. Gwet KL Computing inter-rater reliability and its variance in the presence of high agreement. Br J Math Stat Psychol. 2008 May;61(Pt1):29-48. DOI: 10.1348/000711006X126600.

109. Wongpakaran N, Wongpakaran T, Wedding D, Gwet KL. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples.BMC Med Res Methodol. 2013 Apr 29;13:61. DOI: 10.1186/1471-2288-13-61.

110. Gwet KL. Benchmarking agreement coefficients. [Internet]. 2014 [cited 2019 Feb 18]. Available from: http://inter-rater-reliability.blogspot.com/.

111. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015 Mar 4;10(3):e0118432. DOI: 10.1371/journal.pone.0118432.

112. Pastor-Sanchez JA. Elaboration of controlled vocabularies using SKOS. [Internet]. 2015 [cited 2018 Aug 25]. Available from: http://dcevents.dublincore.org/IntConf/dc-2015/paper/download/404/464.

113. Zeng-Treitler Q, Kim H, Rosemblat G, Keselman A. Can multilingual machine translation help make medical record content more comprehensible to patients? Stud Health Technol Inform. 2010;160(Pt 1):73–7.

114. W3C Wiki. SPARQlEndpoints. [Internet]. [updated 2017 Nov 29; cited 2019 Fev 12]. Available from: https://www.w3.org/wiki/SparqlEndpoints.

115. Stock WG. Concepts and semantic relations in information science. J Assoc Inf Sci Technol. 2010; 61(10):1951–69.

116. Gefen D, Miller J, Armstrong JK, Cornelius FH, Robertson N, Smith-Mclallen A, Taylor JA. Identifying patterns in medical records through Latent Semantic Analysis. Communications of the ACM. Jun 2018: 61(6); 72-7. DOI: 10.1145/ 3209086.

117. Henriksson A1, Moen H, Skeppstedt M, Daudaravičius V, Duneld M. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. J Biomed Semantics. 2014 Feb 5;5(1):6. DOI: 10.1186/2041-1480-5-6.

118. Doing-Harris K, Livnat Y, Meystre S. Automated concept and relationship extraction for the semi-automated ontology management (SEAM) system. J Biomed Semantics. 2015 Apr 2;6:15. DOI: 10.1186/s13326-015-0011-7.

119. Abbas AK, Kumar V, Fausto N. Bases patológicas das doenças. 9rd ed. Rio de Janeiro: Elsevier; 2016. p. 212.

120. de Arruda HF, Costa LF, Amancio DR. Topic segmentation via community detection in complex networks. Chaos: Interdiscip. J. Nonlinear Sci. 2016:26.

121. Gerlach M, Peixoto TP, Altmann EG. A network approach to topic models. Sci Adv. 2018 Jul 18;4(7):eaaq1360. DOI: 10.1126/sciadv.aaq1360.

122. Linked Open Data.[Internet]. [cited 2019 Jul 24]. Available from: https://lod-cloud.net/clouds/lod-cloud.svg.

123. Ngo DH, Truran D, Kemp M, Lawley M, Metke-Jimenez A. Can Wikipedia be used to derive an open clinical terminology? Stud Health Technol Inform. 2019 Aug 8;266:136-141. DOI: 10.3233/SHTI190785.

APÊNDICE 1 – APROVAÇÃO DO COMITÊ DE ÉTICA EM PESQUISA

Aprovado de acordo com parecer consubstanciado do CEP #116123 de 05/10/2012. Mudança de pesquisador aprovada com parecer consubstanciado do CEP # 936.178 de 20/01/2015.

UNIVERSIDADE FEDERAL DE SÃO PAULO - UNIFESP/ HOSPITAL SÃO PAULO

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: CONSTRUÇÃO E VALIDAÇÃO DE UM VOCABULÁRIO DE SAÚDE PARA CONSUMIDORES (VSC) PARA O IDIOMA PORTUGUÊS BRASILEIRO

Pesquisador: Fernando Sousa
Área Temática:
Versão: 1
CAAE: 08548112.6.0000.5505
Instituição Proponente: Universidade Federal de São Paulo - UNIFESP/EPM

DADOS DO PARECER

Número do Parecer: 125.943
Data da Relatoria: 05/10/2012

Apresentação do Projeto:
Consumidores de informação em saúde usualmente têm dificuldade em compreender termos médicos e de saúde devido ao hiato existente entre os vocabulários utilizados por eles e por profissionais de saúde. Este distanciamento é um dos fatores que dificultam, por exemplo, o encontro de informações corretas pelos consumidores quando fazem uma busca sobre saúde na web. Para isso, será proposta a construção e validação de um Vocabulário de Saúde para o Consumidor (VSC) em português brasileiro.

O estudo será composto de três etapas: 1) Preparação dos dados; 2) proposta de um
método para extração automática de termos leigos de saúde (termos candidatos) a partir de fontes heterogêneas da web; e 3) realização de um mapeamento semântico entre os termos candidatos e termos técnicos de vocabulários técnicos de saúde. A validação, executada nas etapas 2 e 3, será realizada junto a profissionais de saúde utilizando a método Delphi. A validação será feita para verificar a validade e relevância tanto dos
termos selecionados, quanto do mapeamento semântico dos mesmos com vocabulários técnicos. Dessa forma, três fontes de dados provenientes da web e escritas majoritariamente por consumidores ou para consumidores de saúde serão avaliadas neste trabalho: 1) notícias sobre saúde provenientes de diferentes fontes; 2) mensagens de usuários no Twitter (mídias sociais); 3) Histórico de buscas de usuários em portais de saúde. Serão selecionados para responder aos

Endereço: Rua Botucatu, 572 1º Andar Conj. 14
Bairro: VILA CLEMENTINO **CEP:** 04.023-061
UF: SP **Município:** SAO PAULO
Telefone: (11)5539-7162 **Fax:** (11)5571-1062 **E-mail:** cepunifesp@epm.br;arpmeleti@unifesp.br

questionários de validação do VSC profissionais de saúde com o seguinte perfil: - Perfil profissional: - Médicos; - Enfermeiros; - Psicólogos; - Fisioterapeutas; - Nutricionistas; - Vínculo público ou privado; - Atendam pacientes de diferentes classes sociais; - Atuem na área acadêmica ou atendam em cínicas; - Sejam formados há 8 anos ou mais.

Objetivo da Pesquisa:

Objetivo Primário:

O objetivo deste trabalho é propor um Vocabulário de Saúde para Consumidor (VSC) baseado em termos provenientes de textos heterogêneas da web.

Objetivo Secundário:

1) Propor um método para a extração automática de termos leigos de saúde a partir de fontes heterogêneas da web, contemplando:¿ A elaboração de uma lista termos candidatos;¿ A validação da lista de termos candidatos.2) Relacionar semanticamente termos candidatos a termos presentes em voca-bulários técnicos de saúde, contemplando:¿ A elaboração do vocabulário propriamente dito;¿ A validação interna do vocabulário.

Avaliação dos Riscos e Benefícios:

Risco mínimo, sem procedimento intervencionista

Comentários e Considerações sobre a Pesquisa:

Pesquisa analisando vocabulário de saúde provenientes de textos da web, e propor um vocabulário para consumidor. Estão descritos os procedimentos do estudo, o qual será conduzido sem financiamento externo, com custo declarado de R$ 17000,00.

Considerações sobre os Termos de apresentação obrigatória:

A folha de rosto encontra-se adequada. Apresenta TCLE contemplando a resolução 196/96

Recomendações:

Sem recomendações

Conclusões ou Pendências e Lista de Inadequações:

O projeto está de acordo com os procedimentos seguidos pelo CEP Unifesp

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

O colegiado acatou o parecer do relator. Projeto aprovado.

UNIVERSIDADE FEDERAL DE SÃO PAULO - UNIFESP/ HOSPITAL SÃO PAULO

SAO PAULO, 19 de Outubro de 2012

Assinador por:
José Osmar Medina Pestana
(Coordenador)

PARECER CONSUBSTANCIADO DO CEP

DADOS DO PROJETO DE PESQUISA

Título da Pesquisa: CONSTRUÇÃO E VALIDAÇÃO DE UM VOCABULÁRIO DE SAÚDE PARA CONSUMIDORES (VSC) PARA O IDIOMA PORTUGUÊS BRASILEIRO

Pesquisador: Fernando Sousa

Área Temática:

Versão: 1

CAAE: 08548112.6.0000.5505

Instituição Proponente: Universidade Federal de São Paulo - UNIFESP/EPM

Patrocinador Principal: Financiamento Próprio

DADOS DA NOTIFICAÇÃO

Tipo de Notificação: Outros

Detalhe: Solicitação de cancelamento enviada indevidamente

Justificativa: O trabalho intitulado "CONSTRUÇÃO E VALIDAÇÃO DE UM VOCABULÁRIO DE SAÚDE

Data do Envio: 06/11/2014

Situação da Notificação: Parecer Consubstanciado Emitido

DADOS DO PARECER

Número do Parecer: 936.178

Data da Relatoria: 20/01/2015

Apresentação da Notificação:

JUSTIFICATIVA DO PESQUISADOR: O trabalho intitulado "CONSTRUÇÃO E VALIDAÇÃO DE UM VOCABULÁRIO DE SAÚDE PARA CONSUMIDORES (VSC) PARA O IDIOMA PORTUGUÊS BRASILEIRO" CAAE: 08548112.6.0000.5505 conduzido por mim, Fernando Sousa, foi cancelado indevidamente. Houve um erro de comunicação, o que resultou num pedido incorreto. Solicito que o pedido seja cancelado de forma que o projeto possa ser conduzido por outro pesquisador.

Objetivo da Notificação:

Solicitação de cancelamento enviada indevidamente e mudança de pesquisador

Avaliação dos Riscos e Benefícios:

na

Endereço: Rua Botucatu, 572 1º Andar Conj. 14
Bairro: VILA CLEMENTINO **CEP:** 04.023-061
UF: SP **Município:** SAO PAULO
Telefone: (11)5539-7162 **Fax:** (11)5571-1062 **E-mail:** cepunifesp@unifesp.br

Comentários e Considerações sobre a Notificação:

Alteração do pesquisador principal para JOSCELI M. TENÓRIO para desenvolver esse projeto, da forma como foi descrito e aprovado.

Considerações sobre os Termos de apresentação obrigatória:

Documentos apresentados e justificados adequadamente

Recomendações:

O CEP-UNIFESP IRÁ REALIZAR A MUDANÇA DO PESQUISADOR PELA PLATAFORMA BRASIL.

Conclusões ou Pendências e Lista de Inadequações:

NOTIFICAÇÃO ACEITA E APROVADA (Solicitação de cancelamento enviada indevidamente e mudança de pesquisador para JOSCELI M. TENÓRIO)

Situação do Parecer:

Aprovado

Necessita Apreciação da CONEP:

Não

Considerações Finais a critério do CEP:

O parecer do relator foi acatado

SAO PAULO, 21 de Janeiro de 2015

Assinado por:
José Osmar Medina Pestana
(Coordenador)

Endereço: Rua Botucatu, 572 1º Andar Conj. 14
Bairro: VILA CLEMENTINO **CEP:** 04.023-061
UF: SP **Município:** SAO PAULO
Telefone: (11)5539-7162 **Fax:** (11)5571-1062 **E-mail:** cepunifesp@unifesp.br

APÊNDICE 2 – TERMO DE CONSENTIMENTO LIVRE E ESCLARECIDO

Termo de Consentimento Livre e Esclarecido (TCLE)

Título:

Validação de termos candidatos para o Vocabulário de Saúde do Consumidor (VSC)

Objetivo e desenho do estudo:

Estudo exploratório para avaliar termos previamente identificados automaticamente de forma a verificar se são referentes à área de saúde, para fins de composição do VSC.

Descrição dos procedimentos:

Preenchimento via web deste questionário. O respondente poderá escolher se deseja continuar avaliando os termos. Não há uma quantidade máxima de termos a serem avaliados. Também poderá retornar à avaliação a qualquer momento. O sistema sorteará os termos a serem avaliados no momento.

Benefícios para o participante:

Trata-se de estudo exploratório, mas a participação possibilitará uma reflexão sobre o uso dos termos de saúde de forma a colaborar na elaboração de um vocabulário.

Garantia de acesso:

Em qualquer etapa do estudo você terá acesso aos investigadores responsáveis pela pesquisa para esclarecimento de eventuais dúvidas. Os principais investigadores estão identificados.

Garantia de liberdade:

É garantida a liberdade da retirada de consentimento a qualquer momento e deixar de participar do estudo, sem qualquer prejuízo.

Direito de confidencialidade:

As informações obtidas serão analisadas em conjunto com as de outros voluntários, não sendo divulgada a identificação de nenhum participante da pesquisa (respondente).

Resultados do estudo:

Também é dado o direito de ser mantido atualizado sobre os resultados parciais da pesquisa, quando em estudos abertos, ou de resultados que sejam do conhecimento dos pesquisadores.

Uso do dado:

É compromisso dos pesquisadores de utilizar os dados e o material coletado somente para este projeto.

Compensações e despesas eventuais:

Este estudo não gera despesas pessoais para o participante em qualquer fase do estudo. Desta forma não será oferecida compensação financeira relacionada à sua participação. Se existir qualquer despesa adicional, esta será absorvida pelo orçamento da pesquisa.

Se você acredita ter sido suficientemente informado a respeito das informações que leu descrevendo o estudo "Validação de termos candidatos para o Vocabulário de Saúde do Consumidor (VSC)"; se ficaram claros para você quais são os propósitos do estudo, os procedimentos a serem realizados, suas garantias de confidencialidade e de esclarecimentos permanentes; e se, também, ficou claro que sua participação é isenta de despesas e concorda voluntariamente em participar deste estudo, podendo retirar o seu consentimento a qualquer momento, antes ou durante o mesmo, sem penalidades ou prejuízo, então por favor preencha sua identificação abaixo para informar aos pesquisadores que obteve de forma apropriada e voluntária o seu consentimento livre e esclarecido para a participação neste estudo, e inicie o questionário.

Responsável:

Josceli M. Tenório, aluna de doutorado do Programa de Pós graduação em Gestão e Informática em Saúde - UNIFESP
<josceli.tenorio@unifesp.br><josceli2006@hotmail.com>

Supervisão:

Ivan Torres Pisa, orientador, Prof. Dr. - DIS - UNIFESP <ivanpisa@gmail.com>

APÊNDICE 3 – IDENTIFICAÇÃO

UNIVERSIDADE FEDERAL DE SÃO PAULO
ESCOLA PAULISTA DE MEDICINA
PROGRAMA DE PÓS-GRADUAÇÃO EM GESTÃO E INFORMÁTICA EM SAÚDE

Pró-reitora:	Profa. Dra. Lia Rita Azeredo Bittencourt Pró-reitoria de Pós-graduação e Pesquisa - UNIFESP
Coordenadora da câmara:	Profa. Dra. Mônica Levy Andersen Câmara de Pós-graduação e Pesquisa da - EPM - UNIFESP
Coordenador do curso:	Prof. Dr. Ivan Torres Pisa, livre docente
Curso:	Mestrado acadêmico
Título do projeto:	Vocabulário de saúde do consumidor em idioma português
Aluno:	Josceli Maria Tenório
Orientação:	Prof. Dr. Ivan Torres Pisa, livre docente
Coorientação:	-
Matrícula:	#51603 – período 01/03/2015 a 28/02/2019
CEP:	Comitê de Ética em Pesquisa da UNIFESP #125.943/12, em 05/10/2012
Linha de pesquisa:	3.Registro, Recuperação e Relacionamento de Dados em Saúde
Grupo de pesquisa:	1.Gestão e Informática em Saúde http://dgp.cnpq.br/dgp/espelhogrupo/6744851350253801 2.Saúde 360 saude360.unifesp.br
Apoio financeiro:	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001

ANEXO 1 – SÍNTESE DOS ESTUDOS RECUPERADOS NA REVISÃO DA LITERATURA

Quadro 4. Síntese dos estudos referentes ao uso, aplicação e/ou desenvolvimento de modelos de CHVs, organizados por data de publicação.

n	Bases	Descrição	Métricas	Achados
1	CID-10	Utiliza métodos de *word embedding** word2vec, Global Vectors, FastText para identificar termos relacionados a um termo "semente" presentes em foruns. O refinamento dos termos selecionados é feito por um método de aprendizagem supervisionada baseado em um CHV existente.[100]	Comparação entre três métodos de *word embedding**. Avaliação dos termos segundo um ranking que indicava o grau de proximidade do termo extraído com um termo médico "semente".	FastText apresentou melhor desempenho. O algoritmo identifica corretamente cerca de 80% dos sinônimos referentes aos top 10 termos de um termo médico.
2	MeSH	Extração de termos a partir de fóruns (8436 questões) e materiais educativos (1428) para criação de um CHV chinês.[97]	Extração manual de termos referentes a doenças relacionados a algumas especialidades médicas, segundo um protocolo.	Foram identificados 1349 termos. Após revisão, 1036 termos do consumidor foram mapeados com 480 termos técnicos.
3	OAC-CHV	Método baseado na similaridade sintática e semântica para identificar os termos do Twitter, preprocessados por meio de word2vec, correspondentes a conceitos de efeitos colaterais (sintomas).[99]	Foram considerados termos válidos os que apresentaram similaridade maior que 0,2.	Para os 10 conceitos de efeitos colaterais mais comuns, identificou 333 termos do Twitter. Apenas 90 são mapeados para o OAC-CHV.
4	Nenhuma	Desenvolvimento de um CHV em idioma francês com termos do domínio de câncer de mama.[22]	Obtenção de termos candidatos por meio de padrões linguísticos, RAT (tf-idf, c-value)	Vocabulário com 192 relacionamentos validados.
5	1. MeSH 2. Wikipedia	Detectar a associação entre problemas de saúde estendendo seu significado e relacionamentos.[98]	Comparação entre métodos automáticos e especialistas humanos na identificação de termos.	O modelo automático pode detectar associações verdadeiras melhor que os especialistas.
6	OAC-CHV	Verificação da cobertura e similaridade léxica dos termos da OAC- CHV para acesso dos pacientes aos relatórios de radiologia.[103]	Cobertura e similaridade léxica dos termos do OAC CHV para radiologia.	Cobertura de 88.5% e similaridade léxica de 72.6% dos termos.
7	OAC-CHV	Desenvolvimento de um sistema para classificar termos leigos minerados de EHR para construir um vocabulário.[101]	Comparação com humanos, precisão e AUC-ROC na identificação de termos.	O sistema mostrou-se efetivo na tarefa de classificar termos, com AUC-ROC > 0.8

n	Bases	Descrição	Métricas	Achados
8	1.UMLS 2.OAC-CHV	Desenvolver um framework (simiTerm) para identificar novos termos do consumidor similares aos existentes na OAC-CHV utilizando perguntas e respostas do Yahoo! Answers.[34]	Medidas referentes a análise de texto.	O sistema apresentou robustez para classificar automaticamente termos similares aos presentes na OAC-CHV.
9	1.OAC CHV 2.UMLS (MetaMap)	Desenvolvimento de um sistema baseado em classificação não supervisionada de termos médicos presentes em relatórios de EHR, utilizando as bases indicadas.[102]	Comparação do sistema a cada um dos métodos citados por meio de AUC-ROC.	AUC-ROC=0.813 Resultado melhor que cada um dos métodos individualmente.
10	1.UMLS 2.MeSH 3.SNOMED 4.CID-9 5. OAC-CHV	Desenvolvimento de um sistema baseado em classificação supervisionada de termos médicos segundo sua importância.[33]	AUC-ROC para comparação do sistema com outros sistemas baseados em NLP.	O sistema recuperou menos termos, porém, a precisão é maior que em outros sistemas (89%) contra 86%.
11	1.OAC CHV 2.UMLS 3.SNOMED	Recuperação de termos de mídia social para atualização da OAC-CHV, restrito aos conceitos referentes a "diabetes".[96]	Cobertura e popularidade dos conceitos nas bases indicadas.	O OAC-CHV apresentou boa cobertura referente ao conceito.
12	1.Behavioral, 2.MedSyn e 3.DBpedia	Método baseado em um classificador supervisionado para mapeamento de sinônimos, comparando a precisão nas bases indicadas.[95]	Mapeamento correto de sinônimos.	Melhoria de 12% em relação às bases indicadas.
13	1.UMLS 2.OAC-CHV	Desenvolvimento de um sistema baseado nos conceitos da UMLS e expressões regulares para verificação da cobertura das bases disponíveis.[104]	Cobertura da terminologia na tarefa de identificar conceitos referentes ao risco de ataque cardíaco.	CHV apresentou a melhor cobertura na tarefa de identificar conceitos referentes ao risco de ataque cardíaco.
14	OAC-CHV	Identificação automática de termos de busca para recuperação da informação em saúde na web.[105]	Compararam três métodos para recuperação da informação	O método baseado no uso do CHV foi mais eficiente.
15	Nenhuma	Mineração de textos de saúde e medicina da Wikipedia para identificação de termos técnicos e do consumidor. A identificação dos termos ocorreu por meio da utilização de frases que denotam uma entre estes.[23]	Os termos técnicos foram identificados segundo sua ocorrência em textos da MedLine, ao passo que os termos do consumidor, a partir de um forum.	Foram identificados 2.721 pares de termos técnicos-consumidor

*Estima a máxima probabilidade de proximidade do termo a outros termos com predição correta.

GLOSSÁRIO

Este glossário foi elaborado exclusivamente para auxiliar a leitura desta tese. As definições foram modificadas por meio de tradução livre do texto das fontes originais e de adequações do texto para melhor atender aos fins.

Conceito Unidade de conhecimento criada por uma combinação única de características. Representações conceituais são artefatos construídos de símbolos.
Um conceito representa um único significado e contém todos os termos que expresse esse significado. Todos os termos dentro de um conceito são sinônimos.

Concept Unique Identifier (CUI) Expressão canônica, ou sequência de letras, números ou símbolos, capaz de identificar de forma única um conceito dentro do recurso terminológico. O identificador de conceito deve ser único dentro do recurso terminológico.

Descritor Os descritores MeSH são os principais termos ou cabeçalhos do vocabulário MeSH usados para descrever os principais tópicos de um item indexado.

Definição Representação de um conceito por uma declaração descritiva que serve para diferenciá-lo de outros conceitos.

Estrutura categórica Utilizada em sistemas de representação de conceitos para assegurar a consistência interna das representações conceituais de composição e o mapeamento com correspondência semântica apropriada entre os conceitos em vários recursos terminológicos.

Hiperonímia Relação que se estabelece entre um vocábulo de sentido mais genérico (*broader*) e outro de sentido mais específico a que pode ser associado.

Hiponímia Subordinação de um termo de sentido mais específico a outro, de sentido mais genérico.

Homonímia 1.Relação entre designações e conceitos em uma dada língua em que uma designação representa dois ou mais conceitos não relacionados.

2.Relação entre dois ou mais vocábulos com som igual mas significado diferente.

Identificador do termo Sequência de letras, números ou símbolos, capaz de identificar de maneira única um termo dentro do recurso terminológico.

Lematização Processo de redução de uma palavra flexionada à sua parte essencial.

Ligação semântica Representação formal de uma relação de equivalência, associativa, partitiva ou hierárquica entre dois conceitos. Termos diferentes podem ter o mesmo significado se forem representações explícitas do mesmo conceito. Isto implica as seguintes características: não redundância, não ambiguidade e não imprecisão.

Meronímia Relação semântica entre os significados de duas palavras, sendo uma delas parte do significado total da outra. Trata-se de uma possibilidade prevista no UMLS para a relação hierárquica entre conceitos.

Monossemia Relação entre designações e conceitos em uma dada língua em que uma designação se refere apenas a um conceito.

N-grama Sequência continua de n palavras em uma sentença.

Palavra Unidade mínima com som e significado que, sozinha, pode constituir um enunciado; vocábulo.

Palavra-chave Palavra que descreve de maneira sucinta e precisa o assunto, ou um aspecto do assunto, discutido em um documento.

Polissemia Relação entre designações e conceitos em uma dada língua em que uma designação representa mais que um conceito compartilhando certas características.

Relação associativa Relação entre dois conceitos que têm uma conexão temática não hierárquica em virtude da experiência. A relação existe sob certo tema de interesse (por exemplo, "uma doença e seu agente causador") e explicitamente reconhecida em virtude da experiência.

Relação Relação entre dois conceitos que podem ser uma relação genérica

hierárquica	ou uma relação partitiva. Ambos podem ser usados para organizar a estrutura dos sistemas terminológicos, mas uma distinção clara dessas duas relações é crítica para a capacidade de fornecer funções de raciocínio. Se um significado mais solto, como “mais amplo que/mais estreito que” for usado, deve ser explicitamente declarado.
Sinonímia	Relação entre os termos em uma dada língua representando o mesmo conceito. Termos que são intercambiáveis em todos os contextos são chamados sinônimos; se eles são intercambiáveis apenas em alguns contextos, eles são chamados quasi-sinônimos.
Stemming	Processo de redução das palavras à sua raiz morfológica.
Stopwords	Palavras mais comuns em um idioma, como preposições. Na computação, *stopwords* são filtradas antes do processamento de linguagem natural.
Termo	1. Uma palavra ou coleção de palavras que compreende uma expressão. 2. Classe de todas as cadeias que são variantes uma da outra, por exemplo, “olho” (singular) e “olhos” (plural) são termos que englobam um mesmo conceito.
Termo candidato	Um n-grama que não é coberto pelo vocabulário de destino, mas pode ser potencialmente adicionado ao vocabulário de destino.
Termo do consumidor	Análogo ao termo leigo. Termos usados por consumidores em saúde para referir um termo técnico ou o termo preferido que represente um conceito.
Termo leigo (*lay term*)	Termos coloquial, usados como sinônimos de termos técnicos.
Termo preferido	Termo principal que representa um determinado conceito.
Termo técnico	Termos preferidos ou descritores presentes em vocabulários controlados ou usados por profissionais.
Terminologia	Representação humana e legível por máquina estruturada de conceitos e relações de saúde.

Tesauro Lista de títulos ou descritores de assuntos geralmente com um sistema de referência cruzada para uso na organização de uma coleção de documentos para referência e recuperação.

Tokenização Processo em que, dada uma sequência de caracteres e uma unidade de documento definida, é a tarefa de separar em unidades, chamados de *tokens*, excluindo caracteres especiais e pontuação.

Vocabulário controlado 1.Representação de objetos em sistemas de organização do conhecimento, incluindo listas, sinônimos, taxonomias e tesauros. Alguns princípios são observados para os termos: sem ambiguidade, controle de sinônimos, identificação de relações semânticas (igualdade, hierárquica, associativa).
2. Sinônimo: tesauro. Lista definida de termos com um significado fixo e inalterável, e da qual uma seleção é feita para a catalogação, resumos e indexação, ou para pesquisa em livros, revistas como assunto e outros documentos. O controle tem a intenção de evitar a dispersão de assuntos relacionados entre descritores diferentes. A lista pode ser alterada ou estendida apenas pelo editor ou agência emissora.

Bibliografia consultada

1. Biblioteca Virtual em Saúde. Descritores em Ciências da Saúde: DeCS. [Internet]. ed. 2017. São Paulo (SP): BIREME / OPAS / OMS. 2017 [updated 2017 May; cited 2015 Jul 28]. Available from: http://decs.bvsalud.org.

2. He Z, Chen Z, Oh S, Hou J, Bian J. Enriching consumer health vocabulary through mining a social Q&A site: A similarity-based approach. J Biomed Inform. 2017 May;69:75-85. DOI: 10.1016/j.jbi.2017.03.016.

3. ISO/FDIS 17117-1. Health informatics -Terminological resources - Part 1: characteristics. 2018.

4. Merriam-webster, Dictionary. [Internet]. 2019 [cited 2019 Jul 31]. Available from: https://www.merriam-webster.com.

5. Michaelis, Dictionary. [Internet]. 2019 [cited 2019 Jul 31]. Available from: http://michaelis.uol.com.br.

6. Mork JG. Building an updated MedLine co-occurrences (MRCOC) file. [Internet]. [updated 2016 Sep 28; cited 2019 Jul 31]. Available from: https://ii.nlm.nih.gov/MRCOC/MRCOC_Doc_2016.pdf.

7. National Labrary of Medicine (U.S.). Unified Medical Language System® (UMLS®). UMLS Glossary. [Internet]. [updated 2012 Aug 15; cited 2018 Jun 10]. Available from: https://www.nlm.nih.gov/research/umls/new_users/glossary.html.

8. Yoo I. Introduction to controlled vocabularies and ontologies. [Internet]. 2012 [cited 2019 Jul 31]. Available from: http://cci.drexel.edu/ieeebibm/bibm12/yoo_2012bibm_tutorial.pdf.