JHI JOURNAL OF HEALTH INFORMATICS

Identificação automática de termos de domínio do consumidor em saúde

Automatic identification of consumer health domain terms

Identificación automática de términos de dominio del consumidor en salud

Josceli Maria Tenório[1], Ivan Torres Pisa[2]

Descritores: Vocabulário controlado; Vocabulário; Informática aplicada à saúde dos consumidores

RESUMO

Objetivo: Deste estudo é descrever um processo de identificação automática de termos candidatos a partir de conteúdos disponíveis na web para fins de construção de um CHV no idioma português-brasileiro. **Método:** Inicialmente realizou-se recuperação de conteúdos da web, especificamente por meio de notícias curtas provenientes de feeds Really Simple Syndication (RSS). Como segunda etapa foram organizados vocabulários de controle baseados em CID-10 e Wikipédia, e finalmente foram aplicadas técnicas de análise de textos para fins de identificação e classificação de termos candidatos. **Resultados:** Foram recuperados 128 termos por meio do vocabulário controlado CID-10. O vocabulário Wikipédia resultou em 411 termos distintos. Os termos foram classificados utilizando a métrica estatística tf-idf possibilitando uma análise de sua relevância. **Conclusão:** A utilização e análise de conteúdos jornalísticos disponíveis na web podem apoiar significativamente o desenvolvimento de um CHV.

Keywords: Vocabulary Crontrolled; Vocabulary controlled; Consumer Health Informatics

ABSTRACT

Objetive: This study is to describe a process for automatic term identification based on web available contents for development of a CHV Brazilian-Portuguese version. **Method:** Firstly we performed a recovery of web contents, specifically using short news coming from Really Simple Syndication (RSS) feeds. In the second step, we organized vocabularies based on ICD-10 and Wikipedia, then we used text analysis techniques to identify and classify candidate terms. **Results:** We identified 128 terms using ICD-10. Using Wikipedia vocabulary we identified 411 unique terms. These terms were classified according to a statistic value tf-idf, revealing the importance of the term. **Conclusion:** Web available contents are able to support the CHV development.

Descriptores: Vocabulario Controlado; Vocabulario; Informática aplicada a la salud de los consumidores

RESUMEN

Objetivo: Este estudio es descubrir un proceso de identificación automática de los términos candidatos mediante contenidos disponibles en web para la construcción de un CHV en el idioma portugués-brasileño. **Método:** Inicialmente se realizó la recuperación de los contenidos web, especialmente por medio de noticias cortas provenientes de feeds Really Simple Syndication (RSS). En una segunda etapa, fueron organizados vocabularios de control, basados en CID-10 y Wikipedia y finalmente se aplicaron técnicas de análisis de textos para fines de identificación y clasificación de términos candidatos. **Resultados:** Se recuperaron 128 términos mediante el vocabulario controlado CID-10. El vocabulario de Wikipedia ha resultado en 411 términos distintos. Los términos fueron clasificados utilizando la métrica estadística tf-idf posibilitando la análisis de su relevancia. **Conclusión:** La utilización y la análisis de contenidos periodísticos disponibles en web pueden apoyar significativamente el desarrollo de un CHV.

[1] Mestre em Ciências. Programa de Pós Graduação em Gestão e Informática em Saúde - PPGIS, Universidade Federal de São Paulo - UNIFESP, São Paulo, (SP), Brasil. Professora do Departamento de Informática e Turismo, Instituto Federal de São Paulo - IFSP, São Paulo, (SP), Brasil.

[2] Professor Livre-Docente do Departamento de Informática em Saúde, Universidade Federal de São Paulo - UNIFESP, São Paulo, (SP), Brasil.

Autor Coorespondente: **Josceli Maria Tenório**
e-mail: **josceli.tenorio@unifesp.br**

Artigo recebido: **09/11/2018**
Aprovado: **27/08/2019**

www.jhi-sbis.saude.ws

INTRODUÇÃO

Consumidores que buscam informação sobre saúde e doenças na web para fins diversos têm sido objeto de estudo desde o início dos anos 2000[1-2]. Questões relacionadas à saúde compõem a segunda área temática mais pesquisada do Google, fornecendo 5% das mais de dois trilhões de pesquisas realizadas em 2016[3]. 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[2]. Isso sugere que os consumidores têm seu foco na utilização de termos e pode ser um caminho potencial para o desenvolvimento de aplicações para o consumidor.

Uma das possibilidades para suportar aplicações desenvolvidas para o consumidor em saúde são os *consumer health vocabulary* (CHV), definidos como vocabulários controlados que ligam palavras cotidianas a termos técnicos[4-5]. Para o idioma inglês existe um CHV bem consolidado, suportado pela National Library of Medicine (NLM) (nlm.nih.gov)[6]. Este vocabulário conta com um mapeamento dos termos voltados para o público em geral com os termos presentes no Unified Medical Language System (UMLS) (nlm.nih.gov/research/umls)[5,7-9]. Diferentemente de outros vocabulários controlados, orientados aos profissionais de saúde, esse CHV possibilita a identificação de um conjunto de termos de domínio do consumidor.

Uma possibilidade de desenvolvimento de CHV considera utilizar conteúdos disponíveis na web e aplicar técnicas de análise de textos para identificação automática de termos[10-11]. Em suma, a análise de texto é composta por duas fases[12]:

1. Preparação de dados: limpeza e pré-processamento, que inclui tokenização, *stemming* e remoção de *stopwords*, além de contagens de termos e tf-idf[13].
2. Análise: métodos de contagem e vocabulário de controle, aprendizado de máquina supervisionado e aprendizado de máquina não supervisionado.

Recentemente um CHV para idioma francês foi construído por meio da aplicação dessas técnicas a conteúdos de um fórum sobre câncer de mama[14]. Métodos baseados em análise de textos recuperados da Wikipédia (wikipedia.org) ou de mídia social têm sido utilizados para fins de atualização do CHV, prospectando novos termos ou sinônimos[15-16].

As contribuições do consumidor para a construção de vocabulários controlados parecem ser seriamente sub pesquisados dentro e fora dos cuidados de saúde[17]. A utilização de bases de dados geradas por consumidores, pode ser um item importante para o propósito de criar um vocabulário de saúde do consumidor.

Este artigo tem como objetivo descrever um processo de identificação automática de termos candidatos a partir de conteúdos disponíveis na web. Trata-se de uma das etapas da construção de um CHV em idioma português-brasileiro[18] que encontra-se em curso no grupo de pesquisa Saúde 360º (saude360.unifesp.br).

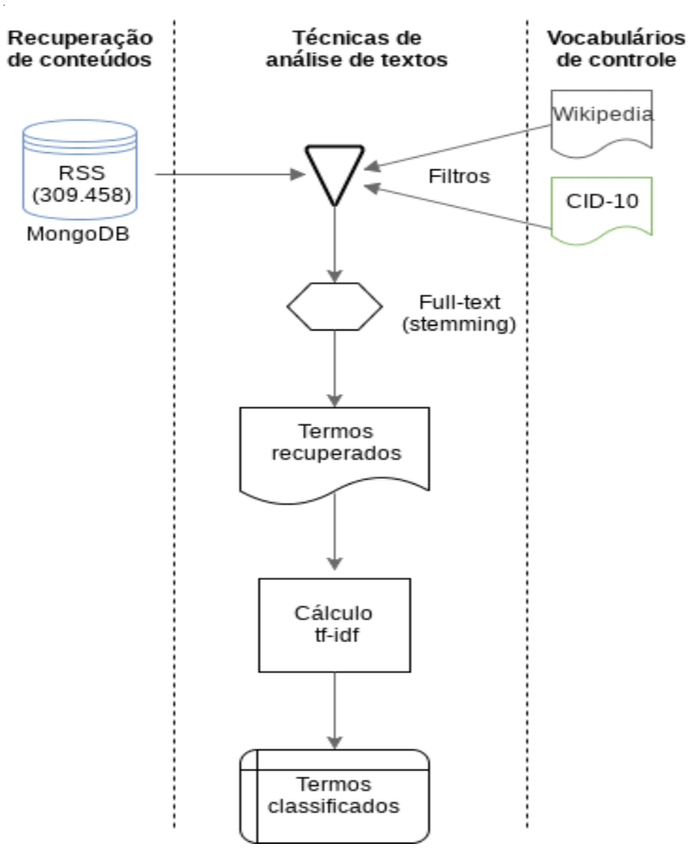

Figura 1 - Etapas metodológicas realizadas no processo de identificação automática de termos candidatos a partir de conteúdos disponíveis na web para apoiar a construção de um CHV em idioma português-brasileiro.

MÉTODOS

O projeto que trata da construção de um CHV em idioma português-brasileiro encontra-se aprovado pelo Comitê de Ética em Pesquisa da Universidade Federal de São Paulo (#936.178/15). Trata-se de um estudo exploratório descritivo com abordagem quantitativa. Para a realização do estudo aqui apresentado foram elaboradas três etapas: recuperação de conteúdos, organização de vocabulários de controle e aplicação de técnicas de análise de textos (Figura 1).

Na etapa 1 foram recuperados conteúdos de ciências e saúde provenientes de 309.458 notícias curtas divulgadas por meio do padrão *Really Simple Syndication* (RSS) publicados entre 2009 e 2017 e obtidos por meio da subscrição de *feeds* como UOL Saúde, G1, Folha de São Paulo, Blog da Saúde e outros. Para efeito de simplificação aqui o termo RSS é usado para referir cada uma dessas notícias. Os conteúdos foram armazenados em uma aplicação para gerenciamento de banco de dados orientada a documento MongoDB versão 3.6. Esta aplicação foi escolhida por possibilitar realizar buscas em textos por termos exatos ou aproximados por meio da aplicação de algoritmos que executam automaticamente o pré-processamento de forma a comparar termos modificados por *stemming*, que realiza a redução do termo ao seu pseudo-lema, em idioma português-brasileiro. O método de busca "*full-text*" do MongoDB usa uma biblioteca de *stemming* chamada *Snowball* (snowballstem.org/). A tarefa de *stemming* realizada por Snowball é baseada no algoritmo de Porter. Desta forma, são tratados plurais e variações de escrita dos termos.

Como etapa 2 para a realização da análise, descrita como uma das fases de análise de textos[12], foram utilizados dois vocabulários de controle: Classificação Internacional de Doenças (CID-10) e Wikipédia (pt.wikipedia.org), um projeto colaborativo escrito por consumidores em saúde. Uma base com 21.424 termos descritivos de doenças do CID-10 foi organizada. Apesar de ser classificado como um vocabulário controlado técnico trata-se de um conjunto importante de termos para verificar a ocorrência em conteúdos com foco no domínio do consumidor. Um vocabulário composto por títulos da Wikipédia referentes às categorias Listas de Medicina (pt.wikipedia.org/wiki/Categoria:Listas_de_medicina) e Medicina (pt.wikipedia.org/wiki/Categoria:Medicina) foi construído. A estas categorias é associado um conjunto de subcategoria que contém os títulos recuperados para compor o vocabulário. Para a categoria "Medicina" foram eleitas as subcategorias com mais de 50 títulos, à exceção de "informática médica", que consta de oito títulos, mas que foi inserido por sua relevância, por exemplo, "cirurgia" e "nutrição". A subcategoria "equipamentos médicos" não estava disponível à época da pesquisa. Para a categoria Listas de Medicina foram recuperados todos os títulos cujo conteúdo disponibilizasse termos referentes à saúde, disponíveis à época da coleta, por exemplo, "bactérias de importância médica (bactérias)" e "doenças raras".

Na etapa 3 de aplicação de técnicas de análise de textos um *script* foi desenvolvido para executar uma busca por termos dos vocabulários de controle CID-10 e Wikipédia na base composta por RSS. Essa tarefa foi realizada por meio da aplicação da técnica de busca *full-text*, suportada por MongoDB. Essa aplicação suporta operações que realizam a busca em textos por meio de *text index* e do operador *$text* (docs.mongodb.com/manual/text-search/). Foi registrada a frequência de ocorrências de cada termo modificado por *stemming* nos RSSs. Para fins de visualização dos termos candidatos foi utilizado o

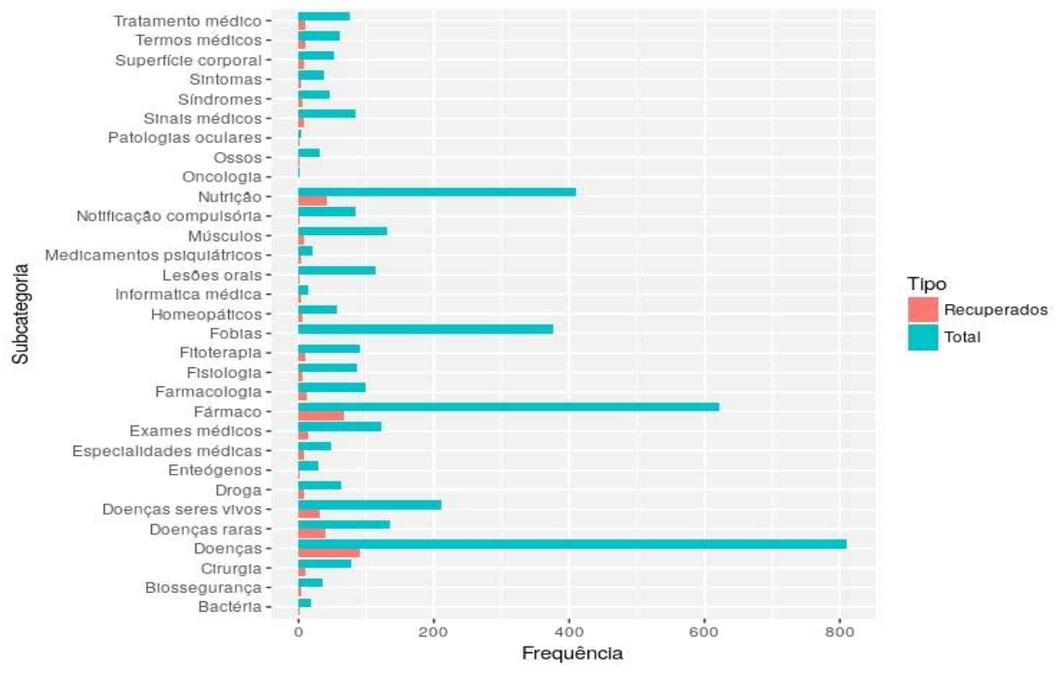

Figura 2 - Frequência de termos componentes do vocabulário Wikipédia em cada subcategoria (Total) e ocorrência de termos recuperados nos RSSs (Recuperados).

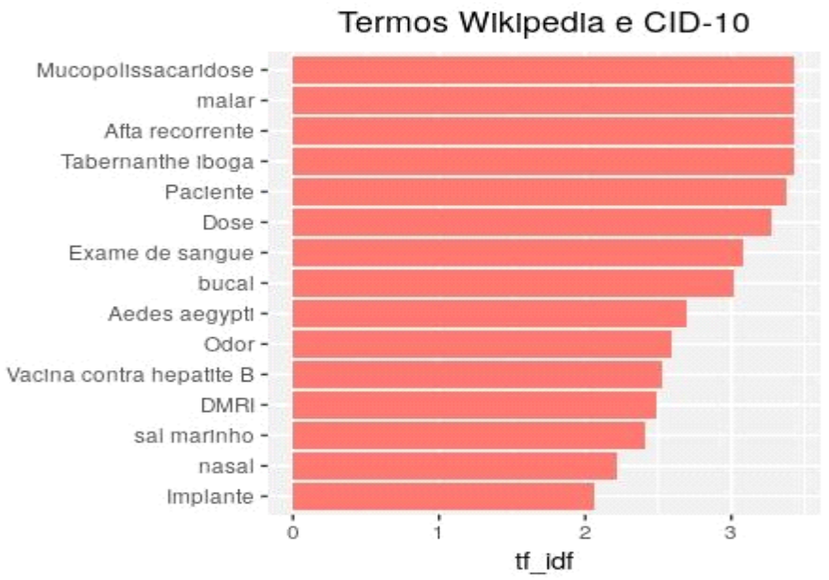

Figura 3 - Primeiros 15 termos de 411 identificados organizados pelo valor de tf-idf.

software R (r-project.org) para execução do cálculo de tf-idf de cada termo. Esse valor está relacionado à relevância do termo nos conjunto de RSS. A construção de vocabulários requer que sinônimos possam ser associados aos termos. Para isso a base de conhecimento DBpedia (pt.dbpedia.org) foi utilizada para recuperar sinônimos mapeados pelos usuários.

RESULTADOS

A Figura 2 mostra a quantidade de termos (títulos) recuperados da Wikipédia para a construção do vocabulário de controle de acordo com as subcategorias referentes a Listas de Medicina e Medicina. O vocabulário Wikipédia foi composto por 4.047 termos. Foram recuperados 128 termos do vocabulário controlado CID-10 referentes a doenças presentes em 2.083 RSSs. A busca por termos do vocabulário Wikipédia resultou em 411 (10,2%) termos distintos presentes em 34.956 RSSs. A relevância de cada termo foi associada ao valor de tf-idf.

A Figura 3 mostra os termos mais relevantes organizados segundo o valor de tf-idf. Observa-se que a maior parte não são referentes a termos comuns, o que pode indicar que o conjunto de termos utilizado para a comunicação do consumidor também contém "termos técnicos".

A Figura 4 mostra os termos mais relevantes, categorizados de acordo com os vocabulários CID-10 e Wikipédia. Em relação à busca por sinônimos na base de dados DBpedia foram mapeados 38 pares "termos preferidos-sinônimos" (Tabela 1).

DISCUSSÃO

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[11]. Neste estudo os esforços referentes ao uso dos termos reais foram contemplados pela construção de um vocabulário baseado em termos utilizados por e para consumidores em saúde. O segundo item foi contemplado pela possibilidade de disponibilizar o conteúdo obtido em formato eletrônico.

Apesar da discussão ainda presente sobre o uso da Wikipédia, no presente estudo os conteúdos recuperados que compuseram o vocabulário foram verificados previamente, o que confere maior credibilidade. Artigos da Wikipédia foram usados para recuperação de termos de saúde e seus sinônimos foi descrita por Vydiswaran.[16] Foram recuperados cerca de 2.700 pares "termo preferido-sinônimo", o que indica que a Wikipédia é uma fonte de termos do consumidor importante. A utilização da Wikipédia como um método para reconhecimento

Tabela 1 - Parte dos pares termo preferidos-sinônimos organizados de acordo com a frequência de ocorrência nos RSSs.

Termo preferido	Sinônimo	Subcategoria	Frequência
Paciente	Pacientes especiais	Termos médicos	19.403
Aedes aegypti	Mosquito da dengue	Doenças seres vivos	3.425
Herpes	Herpes simples	Doenças	242
Odor	Cheiro	Fisiologia	167
Vírus do papiloma humano	HPV	Doenças seres vivos	132
Complexo B	Vitamina B	Fármaco	121
Estirpe	Cepa	Doenças	53
Tumor benigno	Neoplasia benigna	Doenças	52
Anti-hipertensivo	Anti-hipertensor	Doenças	44

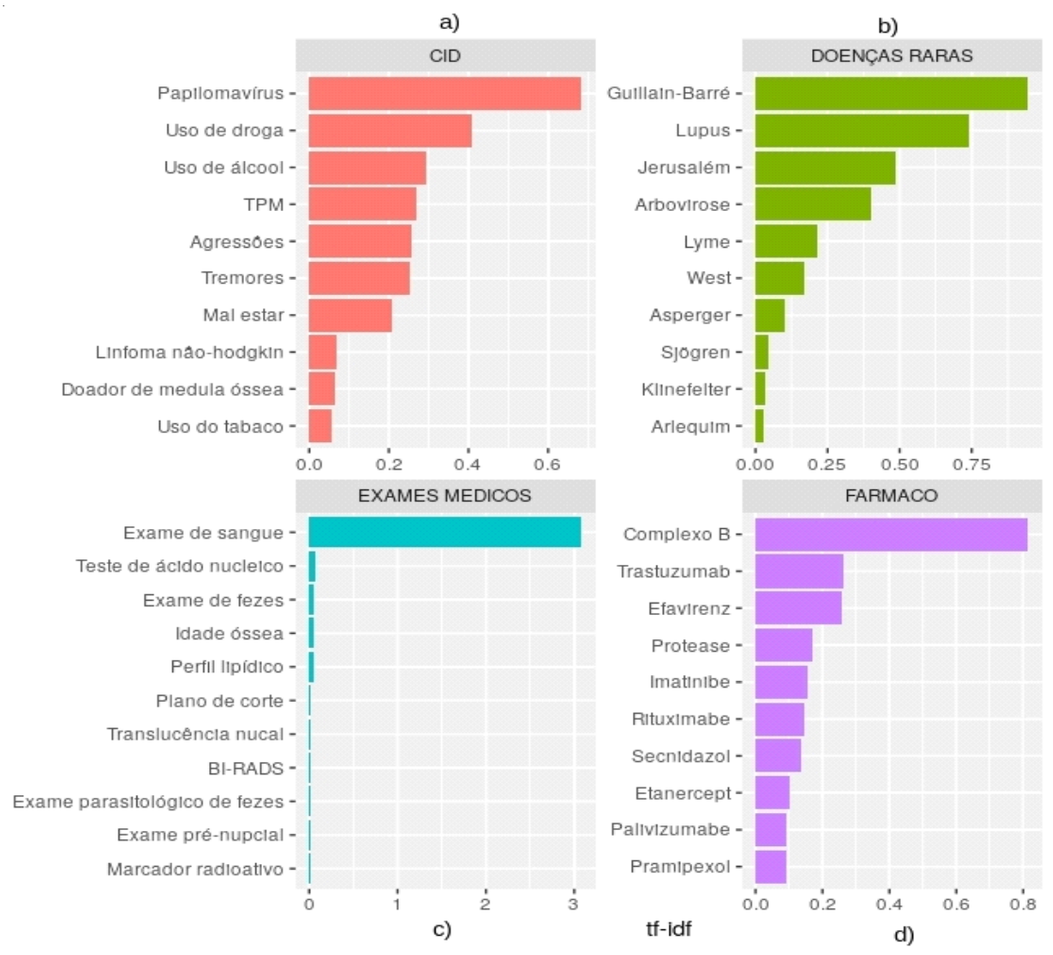

Figura 4 - Gráficos mostrando os termos organizados pelo valor de tf-idf. (a) Termos recuperados por meio do vocabulário CID-10. (b) Termos recuperados por meio do vocabulário Wikipédia pertencentes à subcategorias "doenças raras". (c) Termos recuperados por meio do vocabulário Wikipédia pertencentes às subcategorias "exames médicos". (d) Termos recuperados por meio do vocabulário Wikipédia pertencentes à subcategorias "fármaco".

de termos foi descrita por Astrakhantsev[19], citando diversos exemplos de uso.

Na literatura há estudos que têm descrito a construção ou atualização de CHV por meio da aplicação de técnicas de análise de texto. Porém, os vocabulários de controle utilizados são essencialmente compostos por termos técnicos. Nesse estudo foram utilizados vocabulários diretamente construídos por meio de termos dos consumidores em saúde. Essa escolha propiciou selecionar termos mapeados previamente e utilizados na comunicação com os consumidores. Métodos análogos, baseados na aplicação de técnicas de análise de textos a conteúdos de domínio do consumidor, foram descritos em outros estudos[14-15].

A quantidade de termos de domínio do consumidor está de acordo com outros estudos[14-15]. No estudo de Doing-Harris[15] foram selecionados 774 termos, sendo que 237 foram validados por especialistas. Já em outro foram selecionados 192 termos exclusivos sobre "câncer de mama"[14].

Apesar da aparente diferença significativa entre a quantidade de termos recuperados por meio do vocabulário CID-10 (128) e Wikipédia (411), observa-se que, considerando apenas os termos referentes às subcategorias que versam sobre "doenças", o resultado não contempla essa diferença.

A Figura 4 mostrou que termos caracterizados como técnicos foram frequentemente utilizados na comunicação com o consumidor. Isso indica que os estudos devem ter uma especial atenção no mapeamento de sinônimos e/ou suporte ao consumidor em saúde na utilização desses termos. Trata-se de identificar um conjunto de termos de domínio do consumidor, que é a característica essencial dos CHVs.

Neste estudo o tf-idf foi aplicado apenas como critério para visualização de parte dos termos recuperados nos documentos. Desta forma, foi possível organizar um conjunto de termos (Figuras 3 e 4) e mostrar os mais importantes de acordo com os documentos recuperados. Uma possibilidade a ser verificada por meio do tf-idf seria de que os temos com maior relevância seriam termos comuns. Este estudo mostrou (Figura 3) que isso não é contemplado e que termos menos comuns também fazem parte do rol de termos usados para comunicação com o consumidor.

Uma das limitações desse estudo consta da ausência de uma avaliação quantitativa da acurácia do método, realizada por meio de uma comparação com o padrão-ouro de especialistas. Outro fator importante é que os termos recuperados são dependentes dos vocabulários de controle que compõem o método. Pelo método

termos importantes foram desprezados, porém na continuidade do estudo outras técnicas de análise baseadas em aprendizagem não supervisionada serão aplicadas para fins de recuperação de um número maior de termos potenciais.

CONCLUSÃO

Os resultados obtidos nesse estudo indicam que a utilização de conteúdos disponíveis na web, em conjunto com a aplicação de técnicas de análise de texto, podem colaborar com a identificação de termos para a construção de um CHV na versão português-brasileiro. A base de conhecimento DBpedia mostrou-se importante na tarefa de recuperar sinônimos, apesar da quantidade limitada encontrada. Porém, trata-se de termos mapeados diretamente por consumidores em saúde.

Esse estudo considera um cenário de termos utilizados na comunicação com consumidores em saúde. A essência dos CHV é apoiar aplicações orientadas ao consumidor, como auxiliar na leitura de relatórios médicos, prontuário eletrônico do paciente ou mesmo uma busca na web, promovendo o acesso aos sinônimos ou orientando a exploração do texto. Outra possibilidade consta de disponibilizar serviços análogos a uma versão em idioma português-brasileiro do MetaMap. A expectativa é que essas aplicações favoreçam uma experiência mais efetiva na compreensão dos termos usados.

REFERÊNCIAS

1. Pew Internet and American Life Project. The online health care revolution: how the web helps mericans take better care of themselves [Internet]. 2000 [cited 2018 Mar 4]. Available from: http://www.pewinternet.org
2. 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;324(7337):573-7. doi: https://doi.org/10.1136/bmj.324.7337.573
3. 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 Oct;209(8):342-7.
4. Consumer Health Vocabulary Initiative. [Internet]. [cited 2018 Mar 9]. Available from: http://library.ahima.org/doc?oid=67615#.Wk40XFSdU6g
5. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Inform Assoc. 2006 Jan;13(1):24-9.
6. Unified Medical Language System. 2011AA Consumer Health Vocabulary Source Information [Internet] 2010. [cited 2018 Mar 9]. Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/
7. Unified Medical Language System (UMLS) - Home [Internet]. [cited 2018 Mar 9]. Available from: http://www.nlm.nih.gov/research/umls/
8. 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.
9. 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.
10. Sousa FS, Teixeira F, Nunes FLS, Domenico EBL. Abordagens para construção de vocabulários de saúde para o consumidor: uma revisão da literatura. In: Anais do XIII Congresso Brasileiro de Informática em Saúde - CBIS 2012. Sociedade Brasileira de Informática em Saúde – SBIS. 2012 Nov 19-23. Curitiba,PR; 2012.
11. 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;9(1):e4.
12. 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.
13. 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
14. Tapi Nzali MD, Aze J, Bringay S, Lavergne C, Mollevi C, Optiz T. Reconciliation of patient/doctor vocabulary in a structured resource. Health Inform J. 2018 Jan 1:1460458217751014. doi: 10.1177/1460458217751014.
15. 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.
16. Vydiswaran VG, Mei Q, Hanauer DA, Zheng K. Mining consumer health vocabulary from community-generated text. Proceedings of the AMIA Annu Symp Proc. 2014 Nov 14; Washington DC.1150-9. eCollection 2014.
17. Smith CA. Consumer language, patient language, and thesauri: a review of the literature. J Med Libr Assoc. 2011 Apr; 99(2):135-44.
18. Tenório JM, Torres Pisa I. Consumer health vocabulary: A proposal for a brazilian portuguese language. Stud Health Technol Inform. 2015;216:1089.
19. Astrakhantsev NA, Fedorenko DG, Turdakov DY. Methods for automatic term recognition in domain-specific text collections: a survey. Program Comput Soft. 2015;41(6):336-49. doi: 10.1134/S036176881506002X.