http://journal.unoeste.br/index.php/ca/index
DOI: 10.5747/ce.2019.v11.n2.e278
ISSN on-line 2178-8332

Colloquium Exactarum

Submetido: 28/11/2018 Revisado: 11/02/2019 Aceito: 2/06/2019

UM MODELO PARA EXTRAÇÃO, ESTRUTURAÇÃO, INDEXAÇÃO E RECUPERAÇÃO DE CASOS CLÍNICOS PUBLICADOS NA WEB

A MODEL FOR EXTRACTNG, STRUCTURING, INDEXING AND RECOVERY OF CLINICAL CASES PUBLISHED ON THE WEB

Vitor dos Santos Riedo[1], Silvio Antonio Carro[2]

Universidade do Oeste Paulista–UNOESTE, Presidente Prudente, SP
Faculdade de Informática de Presidente Prudente - FIPP
vitorriedo@gmail.com[1]; silvio@unoeste.br[2]

RESUMO – Atualmente a gama de conteúdos médicos disponíveis na Web é muito vasta, principalmente a de casos clínicos que servem de base para estudos e análises, entretanto muitos desses dados se encontram sem estruturas e originam de fontes heterogêneas, tornando difícil a busca indexação e análise dos dados. Assim esse presente trabalho propõem um modelo de metadados visando padronizar tais conteúdos e servindo como base para utilização de técnicas de mineração de texto convencionais e por PLN para seu povoamento, para futuras buscas, recuperações e manutenções com maior qualidade.
Palavras-chave: Caso Clínico; Mineração de Texto; Mineração de Dados; Processamento de Linguagem Natural; PLN; Metadados; Padrões de Metadados.

ABSTRACT – Currently the range of medical content available on the Web is very wide, mainly clinical cases that serve as the basis for studies and analyzes, however many of these data are without structures and originate from heterogeneous sources, making it difficult to search indexing and data analysis. Thus, this work proposes a metadata model to standardize such contents and serves as a basis for the use of conventional text mining techniques and PLN for their settlement, for future searches, recoveries and maintenance with higher quality.
Keywords: Clinical Case; Text Mining; Data Mining; Natural Language Processing; NLP; Metadata; Metadata Patterns.

1. INTRODUÇÃO

Nos dias de hoje com a constante evolução na área de TI (Tecnologia da Informação), está cada vez mais fácil produzir conteúdo no formato digital, e sua grande maioria estão disponíveis na WEB (World Wide Web). Segundo estudos feitos Gantz e Reinsel (2012) estima-se que em 2020 o volume de dados digitais produzidos no mundo inteiro chegue a 40 zettabyte. Isso devesse principal ao crescimento constante da população, que segundo uma análise bianual do Instituto Francês de Estudos Demográficos INED (2013), estimasse que em 2050 a população mundial chegue a incrível marca de 10 Bilhões de pessoa. Tais conteúdos possuem uma gama gigantesca de assuntos, podendo ser utilizados para diversos fins, como análise, estudos, comparações e outros.

Levando em conta o crescimento populacional, é inevitável que não haja o crescimento no número de pessoas com algum tipo de doença, visto que a cada dia que passa a área médica evolui e novas doenças são constantemente descobertas, com isso a necessidade de saber sobre as mesmas é muito importante.

Sendo assim um dos conteúdos mais valiosos são os casos clínicos, que podem ser usados como base de estudos para adquirir experiência previa de um determinado assunto, sem que seja necessário a vivência do mesmo.

Embora presentes na rede mundial a maior parte das informações vem de locais heterogêneos, semiestruturado ou com nenhuma estrutura, tornando difícil a análise e comparação dos mesmos. Para isso torna-se importante a padronização de tais dados, por meio de um modelo de metadados, pois permite um bom aproveitamento das informações, já que metadados são dados sobre dados e visam modelar, descrever e identificar recursos permitindo buscas mais complexas e precisas, minimizando o tempo de recuperação e manutenção de dados úteis, [3].

O objetivo deste trabalho é apresentar uma solução que possibilite a coleta de casos clínicos publicados em determinados sites WEB, a extração de características importantes, para o povoamento de um modelo de metadados que será proposto, para indexação e recuperações futuras em pesquisas.

E como o intuito é trabalhar com casos clínicos descritos textualmente, é possível a utilização de técnicas exploradas na de Mineração de Texto (Text Mining) para povoamento do metamodelo. Tais técnicas buscam extrair padrões, relações e regras de textos estruturados, semiestruturado ou sem estruturas, assim resultando conhecimento. Bezerra e Goldschmidt (2010).

Para realização do trabalho foi utilizado PLN (Processamento de Linguagem Natural) que trabalha com conteúdo na forma de linguagem natural para geração de informações úteis, com foco na abordagem de EI (Extração de Informação), já que técnica de PLN possui diversos ramos.

A estruturação do artigo está definida como. Na Seção 2 está relatado os trabalhos relacionados, descrevendo os objetivos, metodologias e resultados obtidos. Na Seção 3 será descrito toda a revisão bibliográfica, bem como as metodologias adotadas para desenvolvimento do projeto. Já na Seção 4 é detalhado o projeto Clinical Cases System, a Seção 5 é exposto os resultados obtidos. E por fim na Seção 6 as conclusões.

2. TRABALHOS RELACIONADOS

Na sequência será apresentado trabalhos que propuseram modelos de metadados para padronizar e melhorar a indexação e recuperação de dados.

Carro, Scharcanski e Lima (2013), apresentou um modelo de metadados em RDF (Resource Description Framework), chamado MedISeek, permitindo descrever e recuperar imagens médicas na Web. Seu metamodelo tinha como classe central a imagem, derivando dela classes com dados para melhor detalhamento. Na Fig. 1

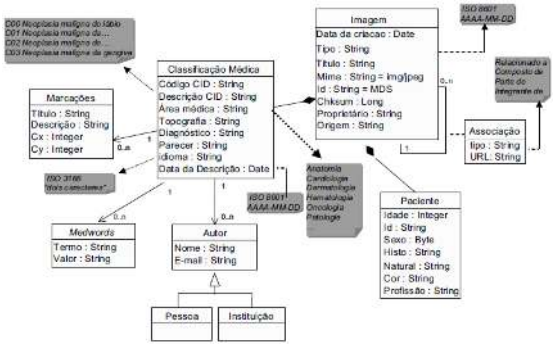
podemos ver o modelo proposto e na Fig. 2 um exemplo prático do mesmo.

Em testes realizados, o modelo de metadados possibilitou uma indexação e recuperação melhores que os resultados obtidos em buscadores Web da época.

Com o mesmo intuito, Brambilla, Carro e Felício (2016), também propuseram um metamodelo baseado na arquitetura Doblin Core, arquitetura essa que visa padronizar recursos digitais, para descrever notícias e matérias jornalísticas, obtendo assim ganhos de recuperação com qualidade superior aos de buscadores da internet.

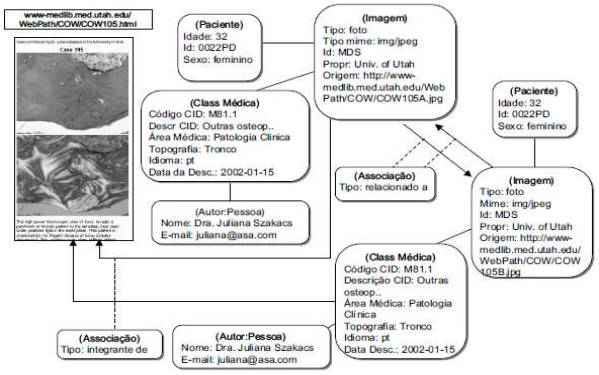
E para povoamento desse metamodelos, Souza (2009) em seu trabalho trabalhou com extração de informação utilizando PLN, mais especificamente utilizando o framework GATE (General Architecture for Text Engineering), realizando comparação entre o modulo ANNIE (A Nearly-New Information Extraction System) e o RELPIE que é proposto pelo trabalho e se baseia em expressões regulares e regras de produção, mostrando que o segundo apresenta bons níveis de extração em textos sem qualquer estrutura, chegando a taxas de 80% em média, e levando em conta os dois resultados, apesar do ANNIE ser mais preciso, entretanto o modelo apresentado obteve melhor cobertura e f-measure. Em 2013 Oleynik (2013) aplicou técnica de PLN na extração de informações em laudos aliado com classificadores bayesianos, obtendo taxas acima de 60% em alguns grupos de informações manipuladas.

Figura 1. Diagrama de classes do modelo MedISeek.



Fonte: Adaptado de Carro, Scharcanski e Valdeni (2003).

Figura 2. Metamodelo povoado.



Fonte: Adaptado de Carro, Scharcanski e Valdeni (2003).

3. REVISÃO BIBLIOGRÁFICA

Esta seção apresenta a fundamentação teórica sobre os métodos utilizados no desenvolvimento deste trabalho, tais como casos clínicos, metadados, mineração de texto e PLN.

3.1. Casos Clínicos

A utilização de casos clínicos para estudos e discussões é uma ótima ferramenta de apoio para os profissionais e estudantes da área médica, e por meio de tal método é possível que eles adquiram conhecimento prévio estudando sem necessidade de viver na prática, obtendo bagagem para futuros casos semelhantes.

Segundo Parente, Oliveira e Celeste (2010), os relatos é o detalhamento dos casos clínicos, onde os mesmos contem características importantes sobre os sinais, sintomas, e outras características do paciente, bem como o relato do diagnóstico, tratamentos, exames e resolução do mesmo.

O estudo desses relatos clínicos se torna ainda mais valiosos quando se trata de casos raros, isso pelo fato de sua pouca ocorrência. Uma doença pode ser definida como rara de diferentes formas, variando de país para país, nos EUA (Estados Unidos da América) a relação é de 1 caso para 200 000 pessoa, já no Japão é a cada 50000 e a Organização Mundial de Saúde (OMS) especifica que para ser rara deve ter uma ocorrência de cerca de 1 caso a cada 2 000 indivíduos, segundo Aronson (2006).

Segundo a National Organization for Rare Disorders (2018) nos EUA o número de doenças raras é de cerca de 7 mil e está quantidade está em constante crescimento [10]. A Fig. 3 representa a fotos de pacientes com a doença Síndrome de Hutchinson-Gilford que é uma doença fatal e atinge 1 criança a cada 4 milhões.

Figura 3. Síndrome de Hutchinson-Gilford.



Fonte: Adaptado de Gonçalves (2014).

3.2. Metadados

A internet possui uma gama vasta de informações, tornando-se um excelente meio de busca e recuperação de conteúdos relevantes, porém se tais dados não possuírem uma documentação adequada fica difícil encontrar conteúdos significante e entendê-los, para isso eles são descritos por meio de metadados, segundo Souza, Catarino e Santos (2012).

A palavra metadados significa dados sobre dados, ele é usado para documentar, possibilitando descrever e identificar recursos das informações para modelar e filtrar seu acesso. Ele se torna extremamente útil na organização e padronização de informações disponíveis na internet, minimizando esforços na recuperação e manutenção de dados uteis, Alves e Souza (2007).

O padrão de metadados DC (Dublin Core) surgiu como uma inciativa de padronizar os recursos na internet, ele é composto por 15 elementos. Sendo caracterizado por sua: flexibilidade, interoperabilidade semântica, simplicidade, consenso internacional e modularidade na Web, de acordo com Campus (2007).

Como a internet possui uma enorme diversidade de dados, ela abrange os dados médicos também, assim se torna indispensável um modelo de metadados para eles. Alguns trabalhos utilizando metadados foram descritos na Seção 2, mas vale ressaltar que Carro, Scharcanski e Lima (2003) desenvolveram o padrão metadados MedISeek para descrever e recuperar imagens médicas, com seus diagnósticos. O padrão possui um grau de detalhamento dos dados muito bom, criando um ótimo modelo para recuperação de informações uteis sobre o caso.

3.3. Mineração de Texto

Segundo Bezerra e Goldschmidt (2010), o processo de mineração de texto visa a extração de padrões, relações e regras de textos estruturados, semiestruturado ou sem estruturas, assim resultando conhecimento.

Para utilizar tal técnica é necessário de início realizar a coleta do material para formar a base de dados para utilização e realizar o pré-processamento.

E como na maioria dos casos esta técnica é aplicada em dados não estruturados é necessário realizar um pré-processamento para criar um padrão para útil ao seu objetivo, Rocha (2017).

Fredigo et al. (2013), cita em seu trabalho 8 técnicas, dentre elas as mais utilizadas são: Filtering, Tokenization, Stemming e Stopword Removal.

Filtering: Visa fazer a remoção de caracteres especiais, de acentuação e pontuação que não alterem o contexto geral do texto. A Tabela 1 representa um exemplo de Filtering.

Tabela 1. Exemplo de filtering

Frase normal	Zico foi o maior jogador da história!
Frase processada	Zico foi o maior jogador da historia

Fonte: O autor.

Tokenization: Separa o conteúdo nas menores unidades de texto possíveis. A Tabela 2 representa um exemplo de Tokenization.

Tabela 2. Exemplo de tokenization.

Frase normal	Zico foi o maior jogador da história!
Frase processada	[Zico] [foi] [o] [maior] [jogador] [da] [história] [!]

Fonte: O autor.

Stemming: Transformas as palavras de volta a sua forma básica, ou seja, para seus radicais (parte da palavra que não muda). A Tabela 3 representa um exemplo de Stemming.

Tabela 3. Exemplo de stemming.

Palavras	Radicais
Pedra	Pedr
Pedreiro	
Pedregulho	

Fonte: O autor.

Stopword Removal: Como o conteúdo que será minerado pode ser muito extenso é possível diminui-lo removendo palavras que não de grande importância e que tem grande aparição, são elas, palavras dos grupos de artigos, preposições, pontuação, conjunções e pronomes, consideradas stopwords, variando essa classificação de acordo com as necessidades. A Tabela IV apresenta uma stoplist que é um conjunto de stopwords.

Tabela 4. Exemplo de stoplist.

A	No	,
O	De	.
Na	;	!

Fonte: O autor.

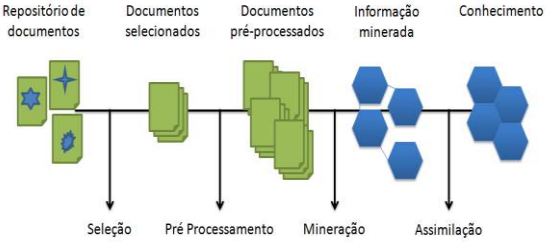
3.4. Processamento de Linguagem Natural

É uma área da IA (Inteligência Artificial) que visa a manipulação e compreensão de textos ou áudios na forma de linguagem natural, possibilitando a geração de informações úteis como tradução e interpretação de textos, extração de informações e outros.

E dentro do PLN temos diversos ramos, porem a técnica que teve foco no trabalho foi a de Extração de Informação (EI) para obter os objetivos.

A EI serve para recuperar determinadas informações de conteúdos principalmente semiestruturados e não estruturados, sendo uma forma de converter para conteúdos estruturados. A Fig. 4 apresenta o fluxo do PLN.

Figura 4. Fluxo PLN.

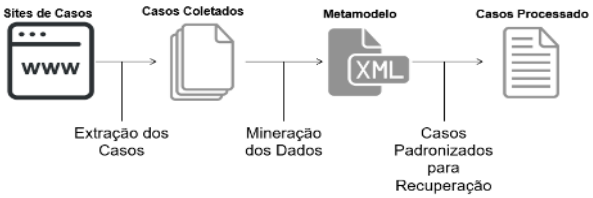


Fonte: https://www.devmedia.com.br/mineracao-de-texto-analise-comparativa-de-algoritmos-revista-sql-magazine-138/34013.

4. PROJETO CLINICAL CASES SYSTEM

Esta seção apresentará o Modelo de Metadados e a arquitetura do sistema Clinical Case System. A arquitetura é composta por 3 módulos: (A) Extração, (B) Mineração e (C) Recuperação. A Fig. 5 ilustra a arquitetura do sistema Clinical Case System.

Figura 5. Arquitetura do sistema Clinical Case System.



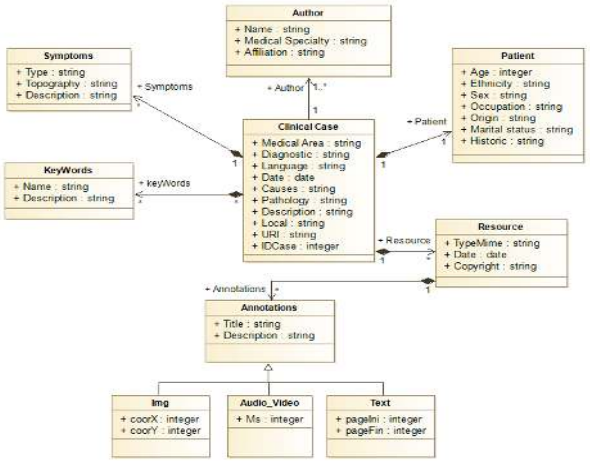
Fonte: O autor.

4.1. Modelo de Metadados Clinical Case System

Para o desenvolvido do metamodelo foi utilizado como base o metamodelo MedISeek de Carro (2003).

Tal modelo de metadados foi criado visando abranger a maior gama de informações úteis para a descrição de um caso clínico. O modelo elaborado é composto por 10 classes, com um total de 35 metadados. A Fig. 6 mostra o Modelo de Metadados Clinical Case System.

Figura 6. Modelo de Metadados Clinical Case System.



Fonte: O autor.

4.2. Extração

A etapa de extração consiste em realizar a coleta dos casos em sites pré-determinados na Web, que vão compor a base de dados do sistema.

Foi utilizado a técnica Wrapper que permite a extração de informações em conteúdos estruturados ou semiestruturados, que foi feita a extração.

Como as fontes de dados se tratava de páginas HTML (HyperText Markup Language) foi possível utilizar tal técnica, entretanto como as mesmas possuíam estruturas heterogêneas, foi necessário desenvolver um modelo para cada fonte de informações.

O primeiro site pré-escolhidos foi o MedPix que tem o seguinte endereço online https://medpix.nlm.nih.gov/home, ele é um site de acesso livre, que contem bases de dados médicas tais como de casos clínicos, imagens médicas e tópicos médicos, e é mantido pelo governo dos Estados Unidos. Já o segundo é LITFL (Life in The Fast Lane) sobre o domínio https://litfl.com que é mantido por médicos que postam diversos conteúdos sobre a área médica.

4.3. Mineração

A mineração foi realizada na base de dados obtida na etapa anterior e aplicada de duas formas diferentes.

A primeira por meio de Wrapper, sendo essa técnica eficaz apenas em dados que possuem algum tipo de estrutura, onde foi desenvolvido um algoritmo próprio para cada fonte por se tratarem de estruturas heterogêneas.

Já a segunda forma de mineração foi utilizando a técnica PLN, foi utilizando a ferramenta CoreNLP[18] a qual possui a funcionalidade de extrair informações nativas como localização, data, pessoas, locais entre outros e também utilizar expressões regulares para combinar os resultados e possibilitar uma extração mais especifica. O grande benefício desse modo de extração é que ele pode trabalhar com dados desestruturados, permitindo que o mesmo algoritmo possa ser utilizado para todos os conteúdos.

Após extraídas as informações, elas irão povoar os metadados do modelo proposto.

4.4. Recuperação

Na etapa de recuperação dos dados foi implementado dois métodos de busca, o primeiro tem como base a busca textual que é submetida a mineração PLN para identificar elementos chaves para busca. Já a segunda busca intitulada como “Busca Avançada” é realizada a pesquisa com filtros.

5. RESULTADOS

Os testes foram realizados sobre uma base de dados composta por mais de 6900 casos clínicos coletados de dois Web Sites diferentes, sendo eles o MedPix e o segundo o LITFL (Life in the Fast Lane) e o objetivo foi povoar 19 dos 35 metadados do modelo. Os 19 metadados são considerados como informações qualificadoras e utilizadas no processo de recuperação. Os demais metadados são informações secundárias e muitas delas subjetivas, mas importantes no relato clínico, portanto devem ser consideradas, mas não foram alvo do processo de povoamento.

A cada caso coletado foi aplicado as duas metodologias de mineração relatadas no capítulo anterior.

A primeira metodologia utilizando wrapper teve que ser preparada individualmente para cada fonte de dados, assim para os casos retirado do site MedPix foi possível coletar 12 metadados e para a fonte LITFL 6 metadados.

Já o segundo método de mineração por PLN, não sofre alterações pela estrutura dos dados, assim o mesmo algoritmo foi utilizado para ambas as fontes, sendo possível extrair até 14 metadados para povoar o metamodelo. A Tabela 5 mostra metadados utilizados em cada método de mineração, sobre cada fonte.

Utilizando a união dos dados que foram passiveis de recuperação obteve-se um resultado de 19 metadados de um total de 35 que compõem o modelo de dados.

E levando em consideração os 19 metadados passiveis de extração pelo sistema, os resultados foram 63.1% e 31.6% referentes ao método por wrapper na fonte MedPix e LITFL respectivamente, e de até 73.7% no modulo por PLN.

Tabela 5. Metadados utilizados em cada métodos de mineração.

Classe	Metadados	Método Wrapper		Método PLN
		LITFL	MedPix	Ambos
Clinical Case	Language	X	X	
	URI	X	X	
	Description	X	X	
	Date	X	X	X
	Diagnostic		X	X
	Medical Area			X
	Causes			X
	Local			X
Patient	Historic	X	X	
	Age		X	X
	Etnicity		X	X
	Sex		X	X
	Occupation			X
	Origin			X
	MaritalStatus			X
Author	Name	X	X	X
	Affiliation		X	X
KeyWords	Name		X	
Symptoms	Name			X
Total		6	12	14
União		19		

Fonte O autor.

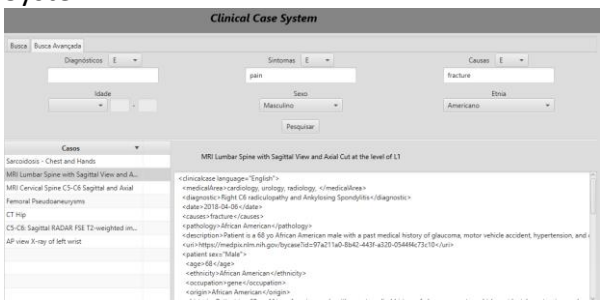
Logo após os testes de povoamento do metamodelo, foi realizado testes de recuperação de casos, em buscadores da internet (Google e Bing) utilizando apenas os sites MedPix (https://medpix.nlm.nih.gov) e LITFL (https://litfl.com/top-100/cxr/) e comparados com módulo de busca proposto na arquitetura do projeto desenvolvido, afim

de testar a precisão de recuperação dos dados.

A primeira observação feita é que o sistema proposto permite realizar buscas por faixa etária, como entre duas idades, maior ou menor que certa idade, já nos buscadores esse comando não é reconhecido, trabalhando apenas com valor especifico e confundindo muitas vezes a idade com um valor qualquer.

Foi realizado então uma pesquisa no sistema por casos de pacientes “Americanos” do sexo “Masculino” que estão com sintomas de “pain” (dor) causado por “fracture” (fratura). A Fig. 7 representa a pesquisa no sistema Clinical Case System.

Figura 7. Pesquisa no sistema Clinical Case System.



Fonte: Oautor.

Nos demais buscadores da web a seguinte frase de pesquisa foi utilizada: “American man with "pain" and "fracture"”, os resultados estão expostos na Tabela 6.

Tabela 6. Resultados primeiro teste.

Buscadores	Resultados	Resultados Precisos	Precisão
Clinical Case System	7	7	100%
Google	9	4	44.4%
Bing	0	0	0%

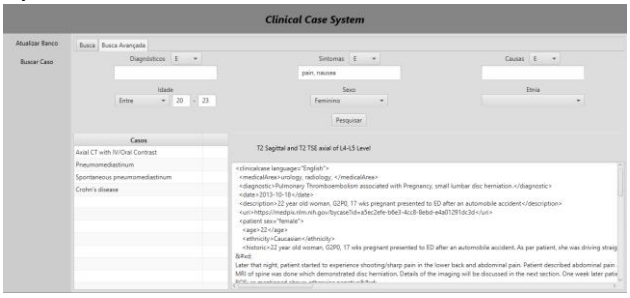
Fonte: O autor.

Após o primeiro teste, foi realizado um segundo visando utilizar principalmente o filtro por idade que o sistema disponibiliza.

Foi realizado a seguinte busca no sistema, mulheres de 20 a 23 anos de idade com dor e náusea. A Fig. 8 representa a pesquisa no sistema Clinical Case System.

O retorno foi com 100% de exatidão na busca feita no sistema e nos buscadores web resultou em bastante erro por confundir os números pesquisados com qualquer número no contexto analisado.

Figura 8. Pesquisa no sistema Clinical Case System.



Fonte: O autor.

Os resultados mostram que como os casos estão estruturados no sistema Clinical Case System e a busca com maior detalhamento se sobressai sobre os demais buscadores, permitindo uma busca com maior rapidez e qualidade nos resultados.

6. CONCLUSÃO

Nesse trabalho foi proposto um metamodelo para padronizar casos clínicos, juntamente com o sistema Clinical Case System que possui módulos de extração de casos clínicos na Web, bem como sua mineração para povoamento do metamodelo e recuperação já no formato padronizado.

Analisando os resultados obtidos na mineração dos casos para estruturação dos mesmos, ficou constatado que o primeiro método é significativamente afetado por mudança no conteúdo que será realizado a mineração, necessitando de constante manutenção, e pela falta de estrutura dos dados, entre tanto se bem configurado e aplicado em dados estruturados ele consegue extrair com perfeição as informações desejadas, desde que as mesmas existam.

Já os resultados obtidos com a segunda metodologia foram mais relevantes

por não necessitar de uma estrutura nos dados minerados.

E com relação ao modulo de recuperação dos dados ficou constatado uma que o metamodelo proposto favoreceu buscas com alta precisão, fornecendo uma melhor base para análise e comparação dos dados.

Pois por meio de comparações de buscas realizadas no sistema e em outros buscadores, o trabalho realizado obteve uma melhor resposta, possibilitando também alguns tipos de buscas que se tornam inviáveis nos buscadores convencionais, por exemplo, a busca por faixa etária de idade.

Para trabalhos futuros fica as seguintes sugestões:

- Implementação de um modulo de cadastramento manual de casos clínicos.
- Aprimoramento no método de mineração por PLN para conseguir abranger uma maior quantidade de metadados e consecutivamente obter melhores resultados.
- Melhoramento do método de extração de casos na Web, visando conseguir uma diminuição na manutenção do mesmo quando ocorre mudanças na fonte.
- Buscar formas de melhorar o tempo de mineração por PLN, que atualmente tem um processamento muito custoso.

REFERÊNCIAS

ALVES, M. D. D. R.; SOUZA, M. I. F. Estudo de correspondência de elementos metadados: DUBLIN CORE e MARC 21. **RDBCI**, v. 4, n. 2,p. 20-38- 2007. https://doi.org/10.20396/rdbci.v4i2.2019

ARONSON, J. Rare diseases, orphan drugs, and orphan diseases. **BMJ**, v.333, 2006. https://doi.org/10.1136/bmj.333.7559.127

CAMPUS, L. F. B . Metadados digitais: revisão bibliográfica da evolução e tendências por meio de categorias funcionais. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação,** v. 12, n. 23, p. 16-46, 2007. doi:https://doi.org/10.5007/1518-2924.2007v12n23p16

BEZERRA, E.; GOLDSCHMIDT, R. A Tarefa de Classificação em Text Mining. **Revista de Sistemas de Informacao da FSMA**,n, 5, p. 42-62, 2010.Disponível em: http://www.fsma.edu.br/si/edicao5/FSMA_SI_2010_1_Tutorial_2.pdf. Acesso em: 18 mar. 2018.

BRAMBILLA, J.; CARRO, S.; FELÍCIO, M. (2016). Descrição e Recuperação de notícias jornalísticas por meio de metadados. **Colloquium Exactarum**, v. 8, n. 1, p. 2016**.** Disponível em: https://doi.org/10.5747/ce.2016.v08.n1.e144. Acesso em: 18 mar. 2018.

CARRO, S. A.; SCHARCANSKI, J.; LIMA, J.V. MedISeek: a web based diffusion system for medical visual information. In: ACM INTERNATIONAL WORKSHOP ON WEB INFORMATION AND DATA MANAGEMENT, 5., 2003. **Proceedings** [...] 2003. p. 54-57). https://doi.org/10.1145/956699.956711

FREDIGO A. H. et al. **Text Mining**. Florianópolis: Universidade Federal de Santa Catarina, 2013. Disponível em: http://www.inf.ufsc.br/~luis.alvares/INE5644/G2_texto.pdf. Acesso em: 15 ago. 2019.

GANTZ, J.; REINSEL, D. **The Digital Universe In 2020:** Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East. IDC iView: IDC Analyze the future, EMC Corporation., 2012.Disponível em: https://www.emc-technology.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf. Acesso em: 18 mar. 2018.

GONÇALVES, A. R. R. **Síndrome de Hutchinson-Gilford ou progéria:** passado, presente e abordagens terapêuticas futuras. 2014. Dissertação (Mestrado) - Universidade do Algarve, Faro, PT, 2014. Disponível em: https://sapientia.ualg.pt/bitstream/10400.1/8352/1/HGPS.pdf. Acesso em: 18 mar. 2018.

INED - INSTITUT NATIONAL D'ÉTUDES DÉMOGRAPHIQUES. **How much inhabitants on Earth tomorrow?**. França, 2013. Disponível em: https://www.ined.fr/en/everything_about_population/demographic-facts-sheets/focus-on/world-population-tomorrow/. Acesso em: 18 mar. 2018.

MANNING, C. *et al.* The Stanford CoreNLP natural language processing toolkit. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: SYSTEM DEMONSTRATIONS. 52., 2014. Proceedings […]. 2014. p. 55-60. https://doi.org/10.3115/v1/P14-5010

NATIONAL ORGANIZATION FOR RARE DISORDERS (Nord). **Rare disease Information**. 2018. Disponível em: https://rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/

OLEYNIK, M. **Extração de informações de narrativas clínicas**. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2013. doi:10.11606/D.45.2013.tde-28112013-185051.

PARENTE, R. C. M.; OLIVEIRA, M. A. P.; CELESTE, R. K. Relatos e série de casos na era da medicina baseada em evidência. **Bras J Video-Sur**, v. 3, n. 2, p. 67-70, 2010. Disponível em: https://s3.amazonaws.com/academia.edu.documents/6743146/bjvs030302_063b.pdf?AWSAccessKeyId=AKIAIWOWYYGZ2Y53UL3A&Expires=1523675005&Signature=xvC88TrNyoHTVk1swwrHXFGg0fc%3D&response-content-disposition=inline%3B%20filename%3DRelatos_e_Serie_de_Casos_na_Era_da_Medic.pdf. Acesso em: 18 mar. 2018.

ROCHA, L. D. H. V. **Mineração de texto aplicada na análise de redações do ENEM**. 2017. Trabalho de Conclusão de Curso (Graduação) – Universidade Federal de Recife, 2017. Disponível em: http://www.bcc.ufrpe.br/sites/www.bcc.ufrpe.br/files/Lucas%20de%20Holanda_0.docx. Acesso em: 18 mar. 2018.

SOUZA, T. B.; CATARINO, M. E.; DOS SANTOS, P. C. Metadados: catalogando dados na Internet. **Transinformação**, v. 9, n. 2, 2012.Disponível em: https://scholar.google.com.br/scholar?hl=pt-BR&as_sdt=0%2C5&q=Metadados%3A+catalogando+dados+na+Internet&btnG=

SOUZA, L. C. **Extração de informação usando integração de componentes de PLN através do framework GATE**. 2009. Dissertação (Mestrado) - Universidade Federal de Pernambuco, Recife, 2009.