Adicionando informações estruturadas ao Bulário Eletrônico da ANVISA

Alternative Title: Adding structured information to the ANVISA's "Bulário Eletrônico"

João Vitor F. da Silva
Programa de Pós-Graduação em Informática
Universidade Tecnológica Federal do Paraná
CEP 86300-000 – Cornélio Procópio – PR – Brasil
poferrari@gmail.com

Carlos N. Silla Jr.
Programa de Pós-Graduação em Informática
Universidade Tecnológica Federal do Paraná
CEP 86300-000 – Cornélio Procópio – PR – Brasil
carlosjunior@utfpr.edu.br

André Y. Kashiwabara
Programa de Pós-Graduação em Informática
Universidade Tecnológica Federal do Paraná
CEP 86300-000 – Cornélio Procópio – PR – Brasil
kashiwabara@utfpr.edu.br

RESUMO

O Ministério da Saúde e outros órgãos relacionados pretendem evitar a automedicação e incentivar o cuidado do uso concomitante entre medicamentos, porém estes órgãos não disponibilizam ferramentas para facilitar este processo. A ANVISA disponibiliza um conjunto de 6.076 bulas em formato PDF, mas as informações nelas contidas não estão estruturadas. Um dos desafios deste trabalho consistiu em extrair automaticamente as informações presentes nesse conjunto de bulas. Este artigo apresenta uma metodologia semiautomática de mineração de textos para mapear as bulas da ANVISA nas redes de interações entre fármacos da base de dados DrugBank, juntamente com as doenças encontradas na base SNOMED-CT. Os medicamentos, as doenças, os fármacos e suas relações foram estruturadas e armazenadas em um banco de dados em grafos utilizando a tecnologia Neo4j.

Palavras-Chave

mineração de textos, bulas, interações, fármaco, doença

ABSTRACT

The Brazilian Ministry of Health and other related organizations are concerned with the issue of self-medication. Although these organizations warn about the risks of concomitantly using different drugs, they do not provide any tools to facilitate this process. ANVISA offers a collection of 6.076 medication guides in PDF file format. However, the information available in this guides are in an unstructured format. One of challenges of this work consisted in the automatic retrieval of information from ANVISAS's medication guides. This paper presents a semiautomatic procedure that maps ANVISAS's medication guides to DrugBank and SNOMED-CT. The medications, the diseases, the drugs, and their relations were structured and stored on a graph database using the Neo4j technology.

SBSI 2015, May 26th-29th, 2015, Goiânia, Goiás, Brazil

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Medical information systems; H.2.8 [**Database Applications**]: Scientific databases

General Terms

Design, Standardization, Human Factors

Keywords

text mining, drug information, interactions, drug, disease

1. INTRODUÇÃO

O tratamento de um paciente doente pode envolver vários médicos com especialidades diferentes e cada médico pode prescrever diversos medicamentos. Desse modo, é importante realizar, de forma científica e racional, a seleção do melhor conjunto de medicamentos considerando prescrições dos médicos de cada especialidade. Em outras palavras, as decisões em relação ao tratamento medicamentoso e as interações estabelecidas entre os médicos são determinantes para o sucesso de um tratamento [5].

No Brasil o tamanho real do problema dos erros de medicação não é conhecido, porém, dados estimados pela Fundação Oswaldo Cruz indicam que 24 mil mortes anuais são ocasionadas por intoxicação medicamentosa [3].

No intuito de contribuir para a tomada de decisão terapêutica, o Ministério da Saúde vem promovendo e incentivando o Uso Racional de Medicamentos. Trata-se de um documento técnico que apresenta uma compilação das Condutas Baseadas em Evidências sobre Medicamentos Utilizados em Atenção Primária à Saúde, constantes no Módulo de Informações do HÓRUS - Sistema Nacional de Gestão da Assistência Farmacêutica [5], que consiste em abordar a lógica da racionalidade na prescrição, dispensação e administração de medicamentos.

Fora a cartilha, existe um sistema desenvolvido que contempla uma coleção de imagens ou arquivos *Portable Document Format* (PDF) das bulas[1] do Ministério da Saúde. Esse sistema não é muito complexo, e conta apenas com uma simples opção de filtros para pesquisa de conteúdo específico do medicamento, indústria farmacêutica, entre outros. O sistema de busca de bulas não possui uma opção de pesquisa simples para verificar quais remédios são indicados para uma determinada doença.

Dentro do nosso conhecimento, não há nenhum trabalho que descreve a utilização das redes de interações para melhorar o processo de pesquisa dos dados disponíveis pela ANVISA. Existem dois trabalhos [13, 11], publicados no começo de 2015, que mostram o interesse da comunidade médica internacional em relação ao estudo de redes de interações entre medicamentos.

O primeiro aborda o problema da utilização de cinco ou mais remédios em idosos e aponta a importância das redes de interações entre drogas e doenças [13]. Este trabalho mostrou que a quantidade de efeitos adversos aumenta de forma não linear a medida que novos medicamentos são adicionados no tratamento [13]. O segundo trabalho apresenta uma metodologia para a construção de uma rede de interações com múltiplos níveis incluindo fármacos, doenças e genes [11] e mostrou algumas propriedades utilizando conceitos da área de redes complexas.

Existem trabalhos que fundamentam o uso de processos de mineração de dados em aplicações voltadas para área de saúde, como o trabalho de Yoon *et al.* [15], em que os autores propuseram um roteiro quantitativo para detecção de reações adversas a medicamentos por meio de registros eletrônicos de saúde dos pacientes de um determinado laboratório.

Já no trabalho de Liu *et al.* [6], é utilizado um algoritmo de mineração de dados para identificar regras de associações entre os medicamentos encontrados em conjunto de registros médicos eletrônicos. Definida essas associações é aplicado um algoritmo para interações entre as regras, sendo possível descobrir falhas na administração concomitante de determinados fármacos.

Um solução próxima a apresentada em Liu *et al.* [6] é encontrada no trabalho de Rho *et al.* [8], em que os autores propõem técnicas de mineração em banco de dados de contraindicações médicas para apresentar regras de associações entre os medicamentos.

Verifica-se que os trabalhos realizados na área pesquisada encontram solução por meio do histórico médico dos pacientes. Contudo, em nenhum destes trabalhos foram utilizadas técnicas para identificar regras de associações entre medicamentos, tendo como base as contraindicações, reações adversas e interações encontradas nas bulas médicas da ANVISA.

No contexto internacional existem alguns aplicativos para o sistema Android que apresentam ferramentas que verificam as interações entre drogas, como o aplicativo Medscape, desenvolvido por WebMD, LLC e o aplicativo Drugs.com Medication Guide, implementado por Drugs.com[2], ambos são muito bem avaliados por seus usuários. No cenário nacional apesar de existirem aplicativos que colaboram com o trabalho dos profissionais da área de saúde. Contudo, desconhecemos a existência de um aplicativo que verifique as interações medicamentosas utilizando as informações das bulas médicas da ANVISA. O nosso trabalho tem um foco regional, ou seja, o sistema será disponibilizado para profissionais que atuam no Brasil.

Neste trabalho, foram utilizadas duas bases de dados: (i) DrugBank[3] [14]; (ii) SNOMED-CT[4] [4].

O DrugBank é uma base de dados que possui informações sobre os fármacos aprovados e não aprovados pelo FDA (*Food and Drug Administration*). O DrugBank também possui um conjunto de interações fármaco-fármaco suportada pela literatura biomédica [14]. Infelizmente, o DrugBank não permite a busca por fármacos associados a doenças utilizando ontologias médicas ou por meio da utilização do CID-10[5] (*Código Internacional de Doenças*).

Já o SNOMED-CT possui a classificação das doenças organizadas utilizando uma ontologia, ou seja, disponibiliza um vocabulário comum da área biomédica organizado em um grafo dirigido acíclico. Contudo, o SNOMED-CT não apresenta os medicamentos que tratam cada uma das doenças.

O DrugBank e o SNOMED-CT fornecem juntos uma importante fonte de informação estruturada para este projeto. Este trabalho propõe a utilização do DrugBank, SNOMED-CT, e as bulas da ANVISA para melhorar o acesso a informações para profissionais brasileiros, permitindo que eles visualizem as redes de interações entre fármacos.

Para realizar esta tarefa, foram utilizadas técnicas de mineração de textos sobre o conjunto do bulário. O sistema implementado realiza a extração semiautomática dos fármacos (princípio ativo e excipientes) e as doenças associadas de cada medicamento. A partir da lista de fármacos identificados para cada medicamento é possível construir a rede de interações medicamento-medicamento por meio da utilização do DrugBank, juntamente com os termos médicos do SNOMED-CT. Note que um medicamento é formado por um ou mais fármacos e é indicado para o tratamento de uma ou mais doenças. Por esse motivo a rede entre os medicamentos é induzida utilizando as interações conhecidas entre fármacos e doenças. Todas essas informações foram armazenadas em um banco de dados baseado em grafo chamado Neo4j.

2. MATERIAIS E MÉTODOS

A Figura 1 apresenta uma visão geral do sistema. O sistema proposto possui quatro etapas principais, são elas: (i) obtenção das bulas do sítio *web* da ANVISA; (ii) preparação, extração e normalização das informações das bulas para cadastro no banco de dados relacional; (iii) integração com outras bases para identificação do conteúdo relevante obtido da normalização do texto das bulas e (iv) inclusão das informações no banco de dados não-relacional desenvolvido por meio do conteúdo relevante encontrado nas bulas, o seu respectivo tópico e nome do medicamento.

2.1 Aquisição dos dados

A ANVISA disponibiliza uma página, denominada Bulário Eletrônico, para a pesquisa de bulas de medicamentos. Na consulta realizada pelos autores deste trabalho em 5 de novembro de 2014, existiam 6.076 bulas disponíveis no Bulá-

[1] $http://www.anvisa.gov.br/datavisa/fila_bula/index.asp$
[2] $http://www.drugs.com/$
[3] $http://www.drugbank.ca/$
[4] $http://www.nlm.nih.gov/snomed/$
[5] $http://www.datasus.gov.br/cid10/v2008/cid10.htm$

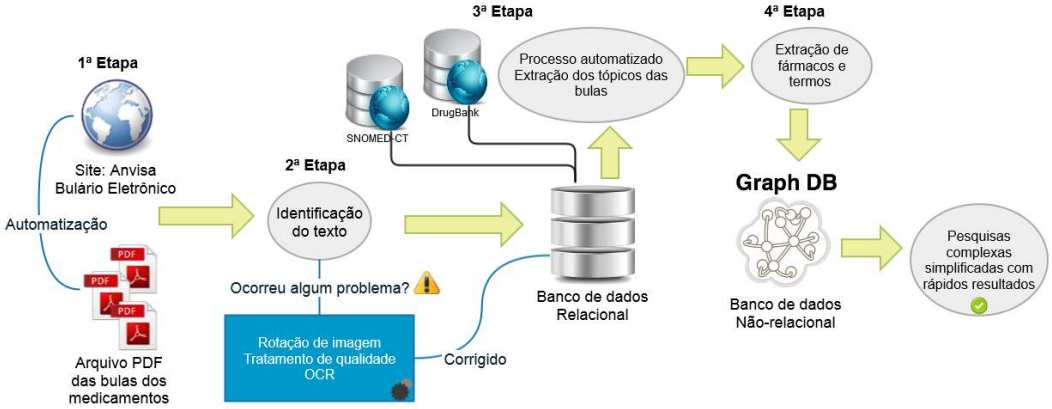

Figura 1: Uma visão geral das etapas que compõem o processo desenvolvido.

rio Eletrônico da ANVISA. O mecanismo de consulta desta página permite utilizar filtros para consultar o nome do medicamente, a empresa fabricante, o número de expediente, o período de publicação e a quantidade de registros por página a serem exibidos.

Para obter as bulas disponíveis no Bulário Eletrônico da ANVISA foi implementado um robô de busca (*web crawler*). O Robô automatizou a obtenção dos arquivos no sistema da ANVISA, realizando o download das bulas médicas por meio de identificação de conteúdo no HTML da página do Bulário Eletrônico.

A sua concepção foi elaborada na plataforma Microsoft Windows, em linguagem C Sharp (`C#`), com a ferramenta IDE Visual Studio 2013 devidamente licenciada para uso. Este robô também foi responsável em capturar o nome do medicamento, a empresa responsável, e outras informações presentes no HTML, seu funcionamento pode ser visualizado pelo vídeo[6] disponibilizado no YouTube.

2.2 Preparação das informações

A etapa de "preparação das informações" consistiu em transformar o arquivo em formato PDF para o arquivo em formato texto. Porém, foram encontrados os seguintes problemas técnicos:

1. Problemas na extração de texto do arquivo PDF, definido pela falta de formatação;
2. Tópicos das bulas estão diferentes do padrão estabelecido pela ANVISA [2];
3. Erros ortográficos no conteúdo da bula;
4. PDF com a imagem da foto da versão física encontrada na caixa do remédio.

Foi desenvolvida uma solução *ad hoc* para cada um desses problemas técnicos. Os pesquisadores interessados podem obter detalhes da implementação através de um contato direto com os autores deste artigo.

O objetivo de solucionar o problema elencado de número 4 é alcançar o maior número de bulas para construção das redes, sendo que apenas 15 destas estavam com a foto física da bula do medicamento.

Desse modo, foi criada uma rotina para recuperar a imagem do PDF, juntá-las e alinhá-las corretamente, além de aplicar um tratamento de qualidade nas imagens, ou seja, tornar a imagem do texto mais nítida. Finalmente, uma ferramenta de *Optical Character Recognition* (OCR) foi utilizada para recuperar o texto. A ferramenta OCR utilizada foi o Tesseract [12].

2.3 Roteiro para segmentação dos tópicos

No arquivo da bula profissional é possível encontrar informações importantes sobre cada medicamento, sendo composta por vários tópicos que auxiliam na prescrição do profissional de saúde. Existem tópicos mais importantes que aparecem com uma certa frequência nos medicamentos e que serão elencados como:

- **Apresentação**: apresenta o medicamento e a empresa responsável por sua elaboração;
- **Composição**: mostra os elementos utilizados para composição do medicamento;
- **Indicação**: para que este medicamento é indicado;
- **Características Farmacológicas**: mostra como o medicamento funciona;
- **Contraindicações**: quando não se pode usar o medicamento;
- **Interações medicamentosas**: o que deve saber antes de usar o medicamento;
- **Posologia e modo de usar**: como deve ser usado o medicamento, onde, como e por quanto tempo pode-se guardar o medicamento;
- **Reações adversas**: quais os males que este medicamento pode causar;
- **Superdose**: o que deve ser feito se usar uma quantidade maior do que a indicada pelo medicamento.

[6] $http://www.youtube.com/watch?v=HdyUJwG9GG0$

A fim de encontrar cada tópico respectivo em cada bula, foi implementado um roteiro, para identificar e marcar a posição de cada tópico, com as seguintes etapas: (i) inicializar a lista de variações de cada tópico; (ii) buscar por expressão regular pela lista de variações; (iii) aplicar um tratamento de texto; e (iv) realizar a marcação do tópico.

Para etapa de pesquisa por expressão regular, foram utilizados os registros da lista de variações, que juntos formavam um padrão de busca no conteúdo da bula, desta forma quando o padrão era identificado aplicava-se um tratamento no texto.

A etapa de "tratamento de texto" tinha a finalidade de retirar os textos desnecessários, deixando apenas o tópico, e não todo seu conteúdo.

Finalmente, utilizando o texto tratado, foi feita a identificação e marcação dos tópicos por meio de busca utilizando expressões regulares.

2.4 Integração com outras bases

Nesta etapa foram utilizada duas outras bases de dados: (i) DrugBank; (ii) SNOMED-CT. Essas bases possuem informações bem estruturadas com relação aos fármacos e às doenças. Desse modo, foram elaborados dois métodos para integração das bulas com outras bases de dados: (1) mapeamento da bula com os fármacos do DrugBank; (2) mapeamento de termos SNOMED-CT com a bula.

Para integrar a informação do DrugBank, foi analisado o texto da composição de cada medicamento manualmente. Os fármacos identificados foram mapeados com seus respectivos `drugbank_id`'s.

Nesse contexto, o mesmo procedimento foi realizado aos textos referentes à indicação, contraindicação e reação adversa das bulas, vinculando-os com as doenças do SNOMED-CT, na qual eram associados aos respectivos `concept_id`'s.

Para realizar a associação entre os termos do Drugbank e do SNOMED-CT (com as bulas da ANVISA) foi necessário realizar a tradução (automática) das informações do DrugBank e do SNOMED-CT.

Após a tradução foi possível cruzar as informações dessas bases com os textos dos tópicos das bulas para encontrar os respectivos `drugbank_id`'s e `concept_id`'s de cada tópico dos medicamentos. Essas informações são então utilizadas para construir a rede de interações entre fármacos, doenças e medicamentos.

2.4.1 *Roteiro para identificação dos fármacos*

No tópico "composição", verificam-se dois tipos de substâncias: (i) princípios ativos; (ii) excipiente. Entende-se pelo princípio ativo o principal fármaco utilizado na composição do medicamento, este responsável pelo efeito farmacológico do remédio, enquanto que excipientes são substâncias utilizadas como veículo para o princípio ativo. É possível encontrar substâncias que são fármacos na lista de produtos excipientes, mas, em geral, essas substâncias (por exemplo, farinha) não são fármacos.

O roteiro de identificação de fármacos foi desenvolvido para analisar o texto da composição do medicamento para identificar aqueles que são princípio ativo ou excipiente.

As etapas que o roteiro realiza para identificação dos fármacos são: (i) leitura do conteúdo da composição do medicamento realizado pela quebra de linha do texto (`\r\n`); (ii) pesquisa por princípio ativo e excipiente determinada por respectivos termos que indicam a qual tipo o termo identificado será vinculado; (iii) tratamento no termo encontrado para evitar retornar palavras que são utilizadas apenas na escrita do texto e não representam um fármaco necessariamente; (iv) procurar o termo identificado nas drogas da base DrugBank, tradução da droga e dos sinônimos para pesquisa nos termos; e (v) criar arquivo texto com o mapeamento dos termos identificados com seus respectivos tipos e código identificador do DrugBank caso seja encontrado.

Figura 2: Exemplo ilustrativo do resultado da identificação dos fármacos no medicamento Lotanol.

A Figura 2 apresenta um exemplo do processo de identificação de fármacos realizada no tópico composição do medicamento Lotanol, na qual elenca-se todas as etapas realizadas pelo algoritmo desenvolvido.

Destaca-se pela Figura 2 a presença do marcador `=[COMP]` em vermelho que indica o início do texto referente à composição da bula, na qual todo o conteúdo é separado por meio das quebras de linhas (`\r\n`).

O roteiro desenvolvido realiza a leitura linha a linha até encontrar palavras que indicam início de um príncipio ativo ou excipiente, que respectivamente estão exemplicados na figura pela palavra "CONTÉM" na cor azul e "EXCIPIENTES" na cor laranja, e limitados pelos caracteres "...." sombreado em amarelo, quando identificado um príncipio ativo, e para excipiente delimitado pelo conjunto de caracteres ":", ",", "\" e "E" apresentada na cor roxa.

Por fim, entende-se pela Figura 2 que os valores sombreados em cinza representam os termos identificados no texto como princípio ativo e excipente, desta maneira exclui-se palavras que não são necessárias para identificação dos fármacos na base DrugBank.

2.4.2 *Roteiro para encontrar os termos médicos referentes às doenças*

Para integrar os termos médicos do SNOMED-CT foi necessário utilizar uma ferramenta de tradução, pois os termos constantes nesta base se encontravam em língua inglesa o que dificultava sua vinculação com o texto as bulas.

Após a tradução dos termos foi realizada uma etapa para tratamento destes na qual o uso de *Stop-words*[7] da língua

[7] *http : //www.ranks.nl/stopwords/portugese*

portuguesa foi responsável pela remoção de palavras desnecessárias de todo o texto. Por fim, tem-se a busca destes termos médicos com os tópicos dos medicamentos, que ao serem encontrados, neste texto, foram mapeados aos respectivos `concept_id`'s.

2.5 Banco de dados baseado em grafos

A informação em relação ao mapeamento dos termos médicos `concept_id` e ao mapeamento dos `drugbank_id` foram armazenadas num banco de dados em grafos. O gerenciador de banco de dados escolhido para esta tarefa foi o Neo4j.

Para visualizar os remédios e suas relações, foi escolhida uma ferramenta disponibilizada pelo próprio Neo4j que apresenta o resultado das consultas em grafo, na qual os vértices, também chamados de nós, podem representar o medicamento, os fármacos do DrugBank e os termos do SNOMED-CT, enquanto as arestas representam as relações entre esses elementos.

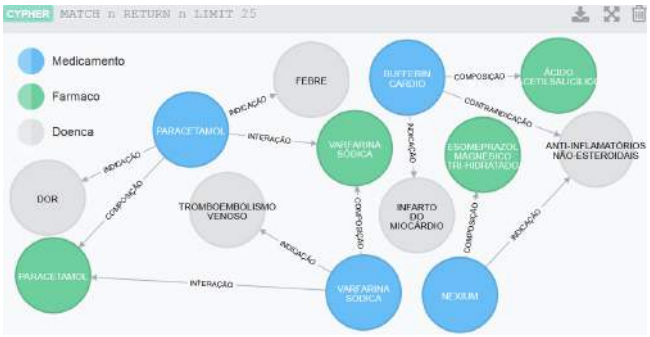

Figura 3: Banco de dados utilizando o Neo4j.

Um remédio é composto por fármacos, e trata um conjunto de doenças (termos SNOMED-CT), além de quais fármacos interagem com outros fármacos. A Figura 3 mostra como as informações entre as bulas estão interligadas e apresenta o retorno de uma consulta no Neo4j que foi desenvolvida pelos dados processados nas etapas anteriores. Para exemplificar o funcionamento do sistema, foi realizada uma consulta sobre o medicamento Paracetamol. O Paracetamol é composto pelo fármaco Paracetamol, mesmo fármaco que possui interação com o medicamento Varfarina Sódica, por este motivo entende-se que o uso concomitante de Paracetamol e Varfarina Sódica possue interações medicamentosas. Essas interações podem trazer a perda de eficácia de algum composto, agravamento de alguma doença, entre outros possíveis problemas [10].

A linguagem de consulta do Neo4j é denominada Cypher e foi inspirada no SQL para descrever padrões em grafos. Ela permite descrever o quê usuário deseja selecionar, inserir, atualizar ou excluir de um banco de dados em grafo sem a necessidade de descrever exatamente como fazê-lo [7].

3. RESULTADOS E DISCUSSÃO

Alguns experimentos preliminares foram realizados com a finalidade de analisar as possíveis soluções para três problemas propostos no presente trabalho: (i) segmentação do texto nos tópicos (indicação, contraindicação, reação adversa, entre outros) definidos pela ANVISA; (ii) identificação dos fármacos em particular reconhecer os princípios ativos e os excipientes utilizados; (iii) mapeamento dos fármacos com o DrugBank. Para que fosse feita a validação de cada processo, foi necessário ter um conjunto de bulas para servirem como referência para fins comparação, ou seja, construir um conjunto *gold standard*. Assim, foram selecionadas de maneira aleatória 100 bulas do total de 6.076, mas eliminando 15 (quinze) bulas que apresentaram problemas no processo de reconhecimento do texto da bula utilizando OCR.

Para cada uma das 100 bulas, foram solucionados de forma manual os problemas propostos, ou seja, os textos das bulas foram segmentados manualmente nos respectivos tópicos. Depois foi realizada a análise das composições de cada medicamento com a finalidade de identificar os princípios ativos e os excipientes para associá-los aos seus respecitvos `drugbank_id`'s.

Os experimentos realizados foram utilizados para validar o roteiro de segmentação e o roteiro de identificação de fármacos. Nas bulas selecionadas para o experimento, o roteiro de segmentação obteve uma precisão média de $89,57\%$, com sensibilidade média de $95,98\%$ e F-score de $92,41\%$. Este resultado é promissor, porém ainda existe uma grande margem para melhorar a precisão.

Outro experimento foi realizado a fim de encontrar os fármacos utilizados no tópico composição de cada bula do medicamento escolhido, por meio dos fármacos presentes na base do DrugBank. A quantidade total de fármacos nas 100 bulas foi de 1.017, das quais 1.017, 185 são princípios ativos. O roteiro desenvolvido reconheceu corretamente um total de 982 fármacos onde 122 eram princípios ativos.

3.1 Segmentação de tópicos

No primeiro experimento, foi realizada a segmentação do conteúdo da bula para realizar a marcação dos diferentes tópicos. Em um primeiro momento, foi desenvolvido um roteiro que converte o texto original de formato PDF em um arquivo de texto puro. Para criar o *gold standard*, utilizando o texto original, foi realizada a leitura e identificação de maneira manual, na qual consistia em ler todo o arquivo de texto e realizar a marcação do tópico respectivo, por padrão foi utilizado o marcador `=[TOPICO]=[início do nome do tópico identificador]` que era adicionado ao texto da bula.

```
genérico Lei nº 9.787, de 1999.

=[TOPICO]=[INDE]APRESENTAÇÕES

Cápsula 150mg
Embalagens contendo 1, 2, 100, 200 e 500 cápsulas.

USO ORAL
USO ADULTO

=[TOPICO]=[COMP]COMPOSIÇÃO
Cada cápsula contém:
fluconazol.....150mg
Excipiente q.s.p.....1 cápsula
Excipientes: álcool etílico, povidona, manitol, celulose microcristalina
e estearato de magnésio.

=[TOPICO]=[INFO]INFORMAÇÕES TÉCNICAS AOS PROFISSIONAIS DE SAÚDE
```

Figura 4: Marcação dos tópicos da bula Fluconazol.

Na Figura 4, exibe-se um exemplo de marcação realizada na bula do medicamento Fluconazol, pode-se visualizar a marcação do tópico *Apresentações* que representa o tópico *Identificação* do medicamento e utiliza o marcador `=[TOPICO]=[INDE]`, juntamente com a marcação do tópico *Composição* que utiliza o marcador `=[TOPICO]=[COMP]`.

Conforme citado no início deste capítulo, a marcação manual foi realizada em 100 arquivos selecionados de maneira aleatória. Pode-se perceber na análise destes que os tópicos eram apresentados de uma maneira sequencial, na qual seguem as normas definidas na Resolução-RDC N^o 47, de 8 de setembro de 2009 da ANVISA, para elaboração e publicação de bulas médicas [2].

A próxima etapa relaciona-se com o roteiro desenvolvido a fim de automatizar a identificação dos tópicos no conteúdo da bula. Os arquivos de texto dos medicamentos foram submetidos ao roteiro que realizava uma busca por expressão regular das variações de cada tópico, quando alguma parte do texto respeitasse o padrão definido da expressão, este era marcado com o respectivo tópico.

Por meio do arquivo marcado manualmente e o outro marcado automaticamente pode-se realizar uma validação no roteiro desenvolvido para identificação dos tópicos das bulas. Para esta tarefa foi implementado um procedimento que recebe o texto marcado manualmente com seu respectivo texto marcado automaticamente, na qual cria uma matriz de confusão com os valores para calcular a precisão (*Precision*), a sensibilidade (*Recall*) e por fim o F-escore (ou *F-measure*).

O experimento foi realizado para todos os remédios selecionados de maneira aleatória, na qual foi realizada uma média entre os valores encontrados pelos métodos de validação, no qual calculou-se como média de precisão o valor de $89,57\%$, juntamente com sensibilidade de $95,98\%$, e por fim *F-score* de $92,41\%$.

3.2 Identificação dos fármacos

Este segundo experimento foi realizado com o objetivo de identificar os fármacos existentes na bula, para o desenvolvimento deste roteiro foi necessária a criação de um arquivo de texto no qual continha apenas as informações do tópico composição, vale ressaltar que esta tarefa foi realizada de maneira automatizada e o conteúdo do tópico composição foi identificado pelo roteiro desenvolvido de marcação.

O mesmo experimento e validações apresentados foram realizados para as 100 bulas selecionadas aleatoriamente, porém o resultado obtido não foi eficaz e necessita de melhorias na identificação de fármacos. Somente o uso da base DrugBank não resolveu o problema por completo da identificação dos fármacos nas bulas, pois alguns termos continuaram sem ser identificados.

Para que os termos pudessem ser identificados com maior facilidade e até melhor organizados foram utilizadas ontologias, na qual tem como principal vantagem a possibilidade de especificar o correto significado e relacionamento entre os termos, evitando interpretações imprecisas sobre o domínio que está sendo modelado [1].

Por meio das ontologias pode-se pesquisar diferentes termos entre as bulas que podem ser sinônimos ou que estão na mesma classe de doença. Além disso, alguns termos podem ser associados a influência sobre uma contraindicação, indicação e até mesmo reação adversa, o que pode ser visto pelo projeto *Disease Ontology* [9]. Na *Disease Ontology* foi criada uma estrutura única para classificação de doenças a fim de unificar a representação da doença entre muitas e variadas terminologias e vocabulários, juntamente com as relações existente entre as doenças.

Como exemplo, imagina-se um determinado medicamento em que sua indicação é prescrita para melhorar a sobrevida após infarto do miocárdio em pacientes clinicamente estáveis. Porém, é contraindicado seu uso concomitante e frequente a um outro que deve ser utilizado com cautela em pessoas com doenças cardiovasculares. Note que "infarto do miocárdio" é uma "doença cardiovascular" e ambos os termos estão relacionados no SNOMED-CT. A contraindicação identificada para termos mais gerais pode ser também utilizada em termos mais específicos [9], melhorando assim a consulta com relação a utilização concomitante entre remédios.

3.3 Exemplos de consultas

Para testar o sistema, algumas perguntas foram elaboradas e as respectivas respostas foram obtidas por meio de consultas feitas utilizando a linguagem Cypher [7]. Por meio do resultado dos testes, pode-se comprovar que as consultas realizadas retornaram as relações entre os medicamentos.

A comparação entre o formato da consulta com a sintaxe SQL de um banco relacional se mostraram de compreensão fácil.

Para ilustrar algumas das funcionalidades do sistema, alguns exemplos estão ilustrados a seguir:

- A seguinte consulta mostra os medicamentos indicados para a doença de Alzheimer:

```
MATCH (med:Bula)-[related]-
(:Doenca {NameDisease: "Alzheimer's disease"})
WHERE Type(related) = "INDICAÇÕES"
RETURN med, Type(related), related LIMIT 20
```

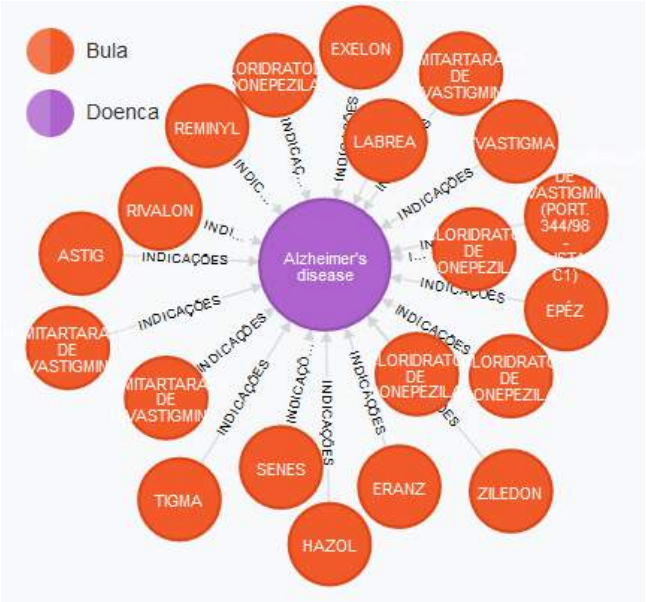

Figura 5: Medicamentos indicados para tratamento de Alzheimer.

A Figura 5 apresenta vinte medicamentos indicados para o tratamento de Alzheimer, que no exemplo elencase por: Exelon, Reminyl, Hazol, entre outros.

- Pela base desenvolvida é possível apresentar hierarquias sobre termos entre as doenças:

```
MATCH p=(Doenca { NameDisease:
    "Alzheimer's disease" })-[r]->(b)
RETURN *
```

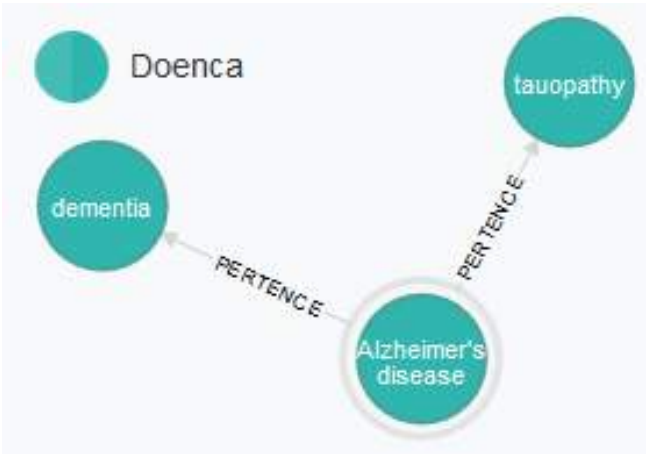

Figura 6: Hierarquias entre as doenças.

A Figura 6 apresenta as hierarquias entre os termos referentes à doença de Alzheimer, pela imagem pode-se visualizar que *Alzheimer's disease* (C26929004) pertence aos termos médicos *tauopathy* (C111479008) e *dementia* (C52448006).

- A consulta a seguir representa os medicamentos que utilizam aspirina em sua composição:

```
MATCH (med:Bula)-[relatedTo:COMPOSIÇÃO]-
(:Droga {DrugBankId: "DB00945"})
RETURN med, Type(relatedTo), relatedTo
```

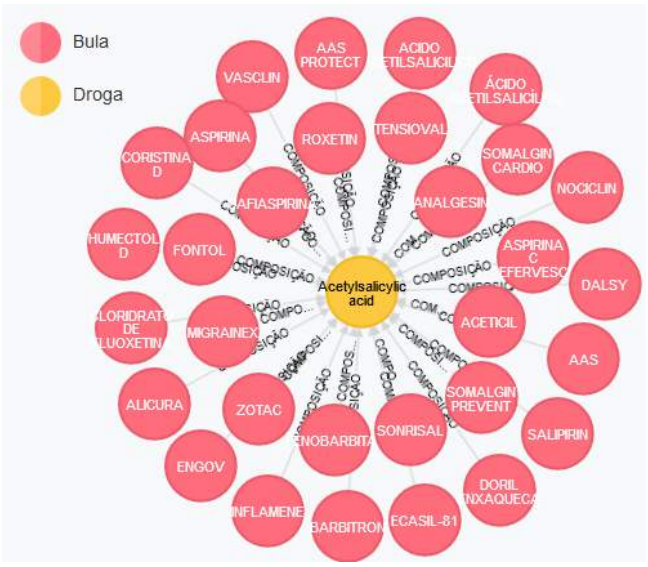

Figura 7: Medicamentos que possuem aspirina em sua composição.

Na Figura 7 é exibido o resultado da consulta realizada utilizando como pesquisa o código do `drugbank_id` ao invés de pesquisar pelo nome do fármaco, que no caso da aspirina seria *Acetylsalicylic acid* (DB00945). Por meio desta imagem percebe-se que a consulta retornou 32 medicamentos que se encaixavam na condição especificada.

- Com relação a consulta anterior, imagina-se que um determinado paciente que possui insuficiência renal necessita tomar um medicamento com aspirina em sua composição:

```
MATCH (dr:Droga {DrugBankId: "DB00945"})<-
 [:COMPOSIÇÃO]-(m:Bula)-[:CONTRAINDICAÇÕES]->
 (do:Doenca {ConceptID: "C192789001"})
RETURN m, dr, do
```

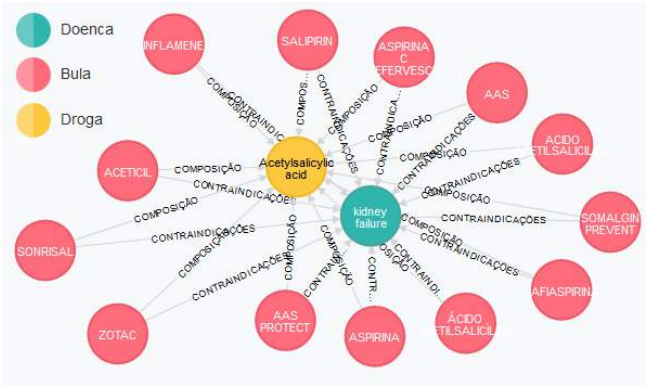

Figura 8: Medicamentos que possuem aspirina em sua composição, porém são contraindicados para insuficiência renal.

A Figura 8 mostra os medicamentos que possuem aspirina em sua composição, mas que são contraindicados para pessoas que tenham insuficiência renal, referente ao termo *kidney failure* (C192789001) da base SNOMED-CT. Comparando-se o retorno apresentado pela atual figura com a Figura 7 visualiza-se que a consulta retornou apenas 13 medicamentos, e não 32 como foi apresentado anteriormente, face a isso, percebe-se que quanto mais condições sejam inferidas, medicamentos mais específicos podem ser encontrados pelas consultas.

A linguagem Cypher possibilita de maneira simplificada diferentes consultas com vários tipos de cruzamento de informações, a estruturação de todas essas informações resulta em uma contribuição para o trabalho dos profissionais brasileiros, retornando as interações entre medicamentos, possíveis contraindicações, medicamentos por doença, entre outras questões.

4. CONCLUSÃO

O presente trabalho propõe uma estrutura baseada em grafo para apresentar as correlações entre fármaco-doença-medicamento, na qual seja de fácil compreensão e visualização. O modelo criado integra os dados não estruturados dos medicamentos obtidos do Bulário Eletrônico da ANVISA, juntamente com as drogas disponibilizadas pelo DrugBank e os termos médicos encontrados na base SNOMED-CT. Por

meio da estruturação dos dados e integração das bases, os resultados apresentados fornecem uma melhoria para a pesquisa de bulas. Este sistema será disponibilizado para profissionais brasileiros que precisam identificar rapidamente as interações medicamentosas, contraindicações e composições.

Nota-se que o uso da tecnologia Neo4j simplifica a complexidade das consultas e retorna informações necessárias para auxiliar no trabalho do profissional da saúde. O sistema de visualização disponibilizada pela ferramenta ajuda na identificação de potenciais efeitos adversos causados pelas interações entre medicamentos.

Como trabalho futuro ainda incidirá no desenvolvimento da integração com outras fontes de informação, por exemplo, consultas através do código CID-10 e *Disease Ontology*.

Além disso pretendemos apresentar o sistema para análise por especialistas, médicos e outros profissionais da área de saúde.

5. REFERÊNCIAS

[1] R. M. d. A. B. J. e. A. d. P. O. A. R. Lamas, J. L. Filho. Ontologias e web services aplicados ao desenvolvimento de sistemas de informação geográfica móveis sensíveis ao contexto. *Anais do V Simpósio Brasileiro de Sistemas de Informaçao (SBSI)*, pages p. 157–168, 2009.

[2] ANVISA. Resolução-rdc n^o 47, de 8 de setembro de 2009, 2009.

[3] S. H. D. B. Cassiani. A segurança do paciente e o paradoxo no uso de medicamentos. *Rev Bras Enferm*, 58(1):95–99, 2005.

[4] R. A. Côté, C. of American Pathologists, A. V. M. Association, et al. *The systematized nomenclature of human and veterinary medicine: SNOMED international*. College of American Pathologists; Schaumburg, IL: American Veterinary Medical Association, 1993.

[5] M. da Saúde. Uso racional de medicamentos: temas selecionados. 1. ed. Brasília: Editora MS, 2012. 156 p.

[6] M. Liu, M. E. Matheny, Y. Hu, and H. Xu. Data mining methodologies for pharmacovigilance. *ACM SIGKDD Explorations Newsletter*, 14(1):35–42, 2012.

[7] Neo4j. Intro to cypher. http://neo4j.com/developer/cypher-query-language/, nov. 2014.

[8] M. J. Rho, S. R. Kim, S. H. Park, K. S. Jang, B. J. Park, and I. Y. Choi. Development common data model for adverse drug signal detection based on multi-center emr systems. In *Proceedings of the 2013 International Conference on Information Science and Applications (ICISA)*, pages 1–7. IEEE, 2013.

[9] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease ontology: a backbone for disease semantic integration. *Nucleic acids research*, 40(D1):D940–D946, 2012.

[10] C. S. Sean and B. Paul. Martindale: the complete drug reference. *Pharmaceutical press1Lamberth High Street, London SEI*, 7:219–599, 2002.

[11] P. G. Sun. The human drug–disease–gene network. *Information Sciences*, 306:70–80, 2015.

[12] Tesseract. Ocr. https://code.google.com/p/tesseract-ocr/, nov. 2014.

[13] J. Wallace and D. S. Paauw. Appropriate prescribing and important drug interactions in older adults. *Medical Clinics of North America*, 99(2):295–310, 2015.

[14] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research*, 36(suppl 1):D901–D906, 2008.

[15] D. Yoon, M. Park, N. Choi, B. Park, J. Kim, and R. Park. Detection of adverse drug reaction signals using an electronic health records database: Comparison of the laboratory extreme abnormality ratio (clear) algorithm. *Clinical Pharmacology & Therapeutics*, 91(3):467–474, 2012.