UNIVERSIDADE FEDERAL DA GRANDE DOURADOS

FACULDADE DE CIÊNCIAS EXATAS E TECNOLOGIA

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

LUCAS SERRANO TULIO

ESTUDO E ANÁLISE DE ASSOCIAÇÕES DE CONCEITOS A PARTIR DE FONTE NÃO ESTRUTURADA

DOURADOS – MS

2016

LUCAS SERRANO TULIO

ESTUDO E ANÁLISE DE ASSOCIAÇÕES DE CONCEITOS A PARTIR DE FONTE NÃO ESTRUTURADA

Trabalho de Conclusão de Curso de graduação apresentado para obtenção do título de Bacharel em Sistemas de Informação pela Faculdade de Ciências Exatas e Tecnologia da Universidade Federal da Grande Dourados.
Orientador: Prof.° Dr. Joinvile Batista Junior

DOURADOS – MS

2016

LUCAS SERRANO TULIO

ESTUDO E ANÁLISE DE ASSOCIAÇÕES DE CONCEITOS A PARTIR DE FONTE NÃO ESTRUTURADA

Trabalho de Conclusão de Curso aprovado como requisito para obtenção do título de Bacharel em Sistemas de Informação na Universidade Federal da Grande Dourados, pela comissão formada por:

_____

Orientador Prof. Dr. Joinvile Batista Junior
FACET – UFGD

_____

Profª. Me. Carla Adriana Barvinski
FACET – UFGD

_____

Prof. Dra. Valguima Victoria Viana Aguiar Odakura
FACET – UFGD

Dourados, 26 de abril de 2016.

RESUMO

Neste trabalho é realizado um estudo de caso a partir de uma fonte majoritariamente não estruturada, abordando o tema de doenças nutricionais, para suportar consultas de relações entre conceitos não explicitamente relacionados no corpus (texto de entrada) utilizado. Inicialmente são definidos os conceitos relevantes para o tema de interesse e então são extraídas manualmente relações binárias de associação entre os conceitos de interesse. Para suportar a consulta de encadeamento entre conceitos é construída uma ontologia na qual são representadas as relações de associação extraídas manualmente do corpus não estruturado, completadas por relações de especialização obtidas na WordNet e por relações de associação ausentes, extraídas de uma fonte adicional. A partir da ontologia criada foi possível automatizar a consulta de encadeamentos de relações de associação, resultando na interligação de conceitos que não aparecem explicitamente relacionados na fonte não estruturada.

Palavras-Chave: ontologia, extração de informação, consultas em ontologias.

ABSTRACT

In this paper is realized a case study with the use of a mostly unstructured source about the nutritional diseases theme, in order to support queries of relations between concepts not explicitly related in the used source. Initially the relevant concepts from the interest theme are chosen, then the binary relations between these concepts are manually extracted. In order to support the queries of chained concepts, an ontology was built in which are represented association relations, manually extracted from the unstructured source, completed by specialization relations from the WordNet and missing association relations from an additional source. The created ontology made possible to automate the queries of chained association relations, resulting in the interconnection of concepts that weren't explicitly associated in the unstructured source.

Keywords: ontology, information extraction, queries in ontologies.

SUMÁRIO

LISTA DE FIGURAS

1. Introdução

Ontologias interligam conceitos relevantes de um determinado domínio através de relações. As relações presentes nas ontologias são extraídas de textos em linguagem natural que abordam temas em um dado domínio de interesse. Ferramentas utilizadas para automatizar a extração de informação ainda apresentam resultados deficientes, extraindo relações incorretas e deixando de extrair relações relevantes. Para evitar estes problemas, foi realizada a extração manual de relações de associação a partir de uma fonte não estruturada (sentenças em linguagem natural).

O principal objetivo deste trabalho é suportar a automação de consultas a partir do encadeamento de relações de associação interligando conceitos. O encadeamento de associações se dá quando os conceitos de uma ontologia estão interligados entre si e formam caminhos que podem ser percorridos automaticamente, a fim de responder a consultas que buscam por conhecimentos não explícitos dentro do domínio. Neste trabalho foi construída uma ontologia no domínio de nutrição, com dados extraídos de fontes não estruturadas, constituída de associações entre os conceitos do domínio de interesse com a finalidade de prototipar o encadeamento de associações.

1.1. Histórico e Motivação

Inferir novos conhecimentos a partir de informações não estruturadas é uma tarefa natural para o ser humano. Requer a capacidade de correlacionar conceitos e extrair informações de dados que, inicialmente, não estão claramente relacionados. Tal tarefa não acontece sem que seja aplicado esforço na interpretação dos dados analisados, e só é possível se, previamente, já se tem conhecimento de determinadas relações dentro do domínio. Isto é, a partir de relações associativas entre conceitos é possível "montar" sequências de associações que relacionam conceitos que, num primeiro momento, não estão conectados. Tais sequencias são definidas por este trabalho como encadeamentos de associação.

O conceito de *Semantic Web* trata da identificação de relações não explícitas por ferramentas que possam compreender documentos e dados semânticos (BENERS-LEE; HENDLER; LASSILA, 2001). Seu objetivo é transformar a *internet* em um meio com maior capacidade de inferência de conhecimentos não explícitos. Para tal, estruturas de representação de conhecimento são utilizadas para representar as

informações do domínio de maneira formal. O conceito de ontologias é um exemplo desse tipo de estrutura (KAPOOR; SHARMA, 2010).

Ontologias são construídas a partir de relações entre conceitos. Relações são propriedades que interligam um conceito a outro em um dado domínio do conhecimento. Relações binárias são associações existentes entre apenas dois conceitos do domínio de interesse. Tais associações interligam os conceitos a fim de criar uma rede de dados a partir da qual podem ser extraídas informações. De maneira geral, uma relação binária é formada seguindo um esquema de "conceito – relação – conceito", na qual não existe um terceiro conceito necessário para a definição da relação. Desse modo, o encadeamento de associações segue um padrão de "conceito – relação – conceito – relação – conceito", envolvendo pelo menos três conceitos. O encadeamento máximo prototipado neste trabalho envolve cinco conceitos.

A extração automática de relações a partir de fontes estruturadas apresenta muitas falhas (TULIO; BATISTA, 2016), omitindo a identificação de associações corretas, e extraindo associações incorretas, o que compromete o resultado do trabalho proposto.

Para contornar esta limitação, a extração manual de relações foi a alternativa escolhida para construir encadeamentos de associações mais robustos que potencializem consultas realizadas sobre o domínio de interesse.

A extração de relações a partir de textos em linguagem natural é uma tarefa que apresenta dificuldades, mesmo quando realizada de maneira manual. Em geral, textos em linguagem natural não são estruturados de forma que as relações estejam expressas claramente. A interpretação do texto é essencial para que as relações relevantes ao domínio em questão sejam extraídas, pois, em linguagem natural, informações frequentemente estão implícitas.

Ao contrário dos textos em linguagem natural, ontologias são caracterizadas por representar conhecimentos de um dado domínio de maneira estruturada, a fim de facilitar a validação e análise das informações. O processo de criação de ontologias acontece quando as informações que estão presentes, inicialmente, em fontes não estruturadas, são extraídas e transformadas em relações binárias. Tais relações são, então, inseridas em ontologias dos domínios de interesse. A forma de representar informações das ontologias permite automatizar a identificação de novos conhecimentos.

A organização estruturada das informações é a base da representação das ontologias que possibilita que relações de associação entre conceitos não originalmente explícitas sejam identificadas automaticamente. Ou seja, o encadeamento entre relações de uma ontologia torna possível a inferência de relações implícitas entre os conceitos do domínio.

1.2. Oportunidades e Relevância

Para os fins desta pesquisa, decidiu-se pela nutrição, especificamente as relações entre nutrientes e doenças, como o domínio de interesse. Tal escolha se dá pela facilidade com a qual é possível entender as relações e os encadeamentos de relações dentro do domínio. É trivial, por exemplo, o conceito de que determinado nutriente pode ser encontrado em algum alimento, ou que o nutriente possua capacidades preventivas ou curativas sobre uma doença. A escolha do domínio de nutrição também se mostra acertada ao se considerar o foco em associações entre conceitos que se encontra nas fontes de informação do domínio, conforme observa-se em Weininger (2015) e no *site* "http://www.healthaliciousness.com".

A fim de criar uma ontologia com relações entre os conceitos do domínio de interesse deste trabalho, foi utilizada uma fonte de informações que pudesse prover relações associando os conceitos selecionados no domínio, que são: *symptom, disease, cause, nutrient* e *food*. Decidiu-se pelo uso da seção da Enciclopédia Britânica que trata sobre doenças nutricionais, escrita por Weininger (2015), que explicita de maneira clara e didática as relações entre doenças nutricionais e os nutrientes cuja deficiência as provoca.

Os cinco conceitos principais selecionados a partir do domínio da nutrição foram escolhidos em virtude da clareza com a qual acontecem as relações associativas entre eles. Em Weininger (2015) percebe-se que as relações de associação são simples de serem identificadas, apesar da complexidade inerente a textos em linguagem natural, que dificultam a extração mecânica das relações sem que haja um tratamento subjetivo utilizando-se de interpretação.

A partir da referência citada foram selecionadas relações de associação entre os conceitos do domínio, as quais foram organizadas como relações binárias a serem incluídas na ontologia. Partindo das frases originais, realizou-se um processo de

interpretação do texto e aquelas relações que se mostraram relevantes ao estudo foram selecionadas.

Uma relação de especialização pode ser explicada como um conceito subordinado a outro. O complexo de vitaminas B, por exemplo, possui diversas vitaminas diferentes (*vitamin B1, vitamin B12, biotin* etc), de modo que uma possível representação de relações de especialização dentro desse grupo poderia ser definida como:

Vitamin

Vitamin B

Vitamin B1

Vitamin B12

Biotin

Relações de associação, por sua vez, interligam dois elementos do domínio dando propriedades e características a esses elementos. De maneira semelhante ao conceito de herança na programação orientada a objetos, elementos descendentes de outros por relações de especialização herdam, obrigatoriamente, todas as relações de associação de seus ascendentes. Um exemplo simples de associação entre conceitos no domínio de nutrição é:

Vitamin C – prevents – Scurvy

O conjunto de relações de associação, quando bem construído e estruturado em uma ontologia, resulta em uma rede de relações entre conceitos que, utilizando-se de ferramentas apropriadas, permite identificar conceitos que não estão explicitamente relacionados nas fontes originais não estruturadas sobre o domínio. As peças (associações) se juntam para formar um todo (encadeamento de associações) a partir do qual novos conhecimentos podem ser extraídos.

1.3. Objetivos do Trabalho

O objetivo geral deste trabalho é definir um processo manual de extração de relações de associação para a construção de uma ontologia, a fim de realizar um estudo de caso do encadeamento de associações do domínio de interesse.

Os objetivos específicos são:

- Avaliar o potencial da caracterização de vários comprimentos de encadeamento de associações, criando uma base para o suporte a consultas no domínio de interesse. O comprimento do encadeamento de associações é definido pelo número de conceitos envolvidos, interligados a partir de relações binárias.
- Construir uma pequena ontologia com as relações extraídas para a realização de um estudo de caso.
- Prototipar encadeamentos entre relações de associação para gerar alternativas de encadeamentos entre conceitos no domínio de interesse.

1.4. Metodologia Adotada

Diversos trabalhos tratam de métodos e metodologias para a construção de ontologias. Trabalhos como os de Iqbal et al. (2013), Corcho, Fernández-López e Gómez-Pérez (2013) e Öhrgren e Sandkuhl (2005) dispõem sobre as diferentes metodologias existentes.

A metodologia utilizada por este trabalho foi construída de acordo com as necessidades encontradas durante o processo de desenvolvimento da pesquisa. No entanto, ela é validada pela literatura da área conforme é possível observar pela análise de trabalhos que dizem respeito à metodologias na construção de ontologias.

De maneira geral, o primeiro ponto de grande parte das metodologias pesquisadas é a definição do domínio, escopo ou objetivo da ontologia. As pesquisas de Noy e McGuinnes (2001), Férnandez, Gómez-Pérez e Juristo (1997), Uschold e King (1995) e Staab (et al, 2001) têm em comum essa característica. Sugumaran e Storey (2002) especificam, ainda, que no primeiro passo da construção de uma ontologia também devem ser identificados os termos básicos do domínio de interesse.

A metodologia proposta por Uschold e King (1995) indica quatro passos que entram de acordo com o trabalho realizado nesta pesquisa. Os passos são: identificar o propósito da ontologia, construí-la, avaliá-la e documentá-la. A presente pesquisa tem,

como objetivo, construir uma ontologia manualmente a fim de avaliar o potencial de pesquisas sobre o encadeamentos de associação.

O processo de construção da ontologia deste trabalho também apresenta semelhanças com o que é definido pela metodologia METHONTOLOGY, desenvolvida por Fernández, Goméz-Pérez e Juristo (1997). Öhrgren e Sandkuhl (2005) resumem a METHONTOLOGY da seguinte maneira: inicialmente, identifica-se o propósito e o escopo da ontologia; a seguir, o conhecimento deve ser coletado, processo que pode acontecer de diversas maneiras, incluindo a análise formal de textos. São selecionados os termos relevantes ao domínio de interesse, e estes termos, ou conceitos, são agrupados de acordo com seus significados. A metodologia trata também da pesquisa de ontologias já existentes que podem ser reutilizadas, e finaliza com a implementação da ontologia, seguida do processo de documentação.

A metodologia iterativa de construção de ontologias proposta por Noy e McGuinness (2001) também apresenta conceitos similares aos utilizados por este trabalho. A pesquisa indica que, após a definição do escopo e domínio da ontologia, devem ser definidos termos importantes a serem utilizados, os quais são organizados hierarquicamente. A seguir, com a hierarquia definida, são incluídas as propriedades, ou relações, entre os conceitos, ou classes, da ontologia. A construção da ontologia se dá de maneira iterativa, de modo que é constante a revisão e a inclusão de novos detalhes a medida que o trabalho é desenvolvido.

A metodologia utilizada por este trabalho segue as seguintes etapas:

- **Primeira etapa:** identificação do domínio de interesse e dos conceitos principais. Seleção da fonte não estruturada de informações sobre o domínio.
- **Segunda etapa:** caracterização de hierarquias de especialização entre conceitos relevantes, utilizando-se da WordNet para caracterizar as relações de especialização.
- **Terceira etapa:** Extração manual das relações de associação entre os conceitos de interesse a partir da fonte não estruturada selecionada. Utilização de uma fonte adicional para complementar associações inexistentes na fonte não estruturada.
- **Quarta etapa:** Agrupamento de relações de associação com base na semânticas das relações e escolha do nome mais relevante para cada grupo de associações. Escolha dos nomes das relações inversas às selecionadas.
- **Quinta etapa:** Construção de uma ontologia a partir das relações extraídas.

- **Sexta etapa:** Análise dos encadeamentos de associações obtidos a partir da ontologia construída.

1.5. Conteúdo do Trabalho

Este trabalho está organizado em quatro seções, organizadas da seguinte maneira:

- O capítulo 1 introduz os conceitos que serão tratados pelo trabalho, indicando também as motivações, os objetivos e a relevância desta pesquisa.
- O capítulo 2 caracteriza a fundamentação teórica, que dispõe sobre o conceito de ontologias e sobre a WordNet.
- O capítulo 3 detalha os processos executados para o desenvolvimento do projeto.
- O capítulo 4 apresenta os resultados alcançados e propõe trabalhos futuros.

2. Fundamentação Teórica

Nesta seção são apresentados: a WordNet e o conceito de ontologias.

2.1. A WordNet

A WordNet é um banco de dados léxico composto de conceitos de substantivos, verbos, adjetivos e advérbios, os quais são interligados para que se forme uma rede de palavras e conceitos (PRINCETON UNIVERSITY, 2010). Na WordNet os conceitos são apresentados utilizando-se de uma estrutura hierárquica, organizando-os de maneira similar ao que acontece em ontologias, que também utilizam relações de especialização na organização dos conceitos.

Os conceitos apresentados pela WordNet contém, conforme exemplificado pela Figura 1, termos sinônimos e breves textos explicativos que esclarecem a significado do conceito. A Figura 1 também apresenta uma amostra da organização em hierarquia com a qual a WordNet organiza as informações.

vitamin -- (any of a group of organic substances essential in small quantities to normal metabolism)
=> fat-soluble vitamin -- (any vitamin that is soluble in fats)
=> vitamin A, antiophthalmic factor, axerophthol, A -- (any of several fat-soluble vitamins essential for normal vision; prevents night blindness or inflammation or dryness of the eyes)
=> vitamin A1, retinol -- (an unsaturated alcohol that occurs in marine fish-liver oils and is synthesized biologically from carotene)
=> vitamin A2, dehydroretinol -- (a viscous alcohol that is less active in mammals than is vitamin A1)
=> vitamin D, calciferol, viosterol, ergocalciferol, cholecarciferol, D -- (a fat-soluble vitamin that prevents rickets)
=> vitamin E, tocopherol, E -- (a fat-soluble vitamin that is essential for normal reproduction; an important antioxidant that neutralizes free radicals in the body)
=> alpha-tocopheral -- (a potent form of vitamin E obtained from germ oils or by synthesis)
=> vitamin K, naphthoquinone, antihemorrhagic factor -- (a fat-soluble vitamin that helps in the clotting of blood)
=> vitamin K1, phylloquinone, phytonadione -- (a form of vitamin K)
=> vitamin K3, menadione -- (a form of vitamin K)
=> water-soluble vitamin -- (any vitamin that is soluble in water)
=> B-complex vitamin, B complex, vitamin B complex, vitamin B, B vitamin, B -- (originally thought to be a single vitamin but now separated into several B vitamins)
=> choline -- (a B-complex vitamin that is a constituent of lecithin; essential in the metabolism of fat)

Figura 1: Textos da WordNet para a hierarquia de especialização de *Vitamin* (retirado de Miller (1995)).

A WordNet difere de um dicionário comum pois neste último a definição dos conceitos se dá de maneira que não são apresentados as relações do conceito em questão, sejam estas de associação, especialização, generalização ou agregação. A WordNet organiza seus conceitos interligando-os com outros conceitos que agregam à

definição e ao significado do que se procura. O conceito de *tree* (árvore), por exemplo, é definido por um dicionário com uma frase que atribui sentido à palavra. A WordNet, além da definição, apresenta também quais conceitos, ou partes, formam uma árvore, que tipos de árvores existem e a partir de quais conceitos derivam o conceito de árvore (MILLER, 1993).

A WordNet apresenta mais semelhanças com um *thesaurus* (dicionário de sinônimos) do que com um dicionário comum. Isso se dá pois a WordNet organiza seus conceitos com base no significado das palavras, e não com base em suas formas (MILLER et al., 1993). Isso significa que a pesquisa de uma palavra elenca todos os seus sentidos e os sinônimos existentes para cada sentido. A Figura 2 apresenta o resultado obtido com a pesquisa da palavra *word*. Observa-se que existem dez resultados para o substantivo *word* e um resultado para o verbo *word*. Para cada significado também são listados, quando cabível, os sinônimos do conceito.

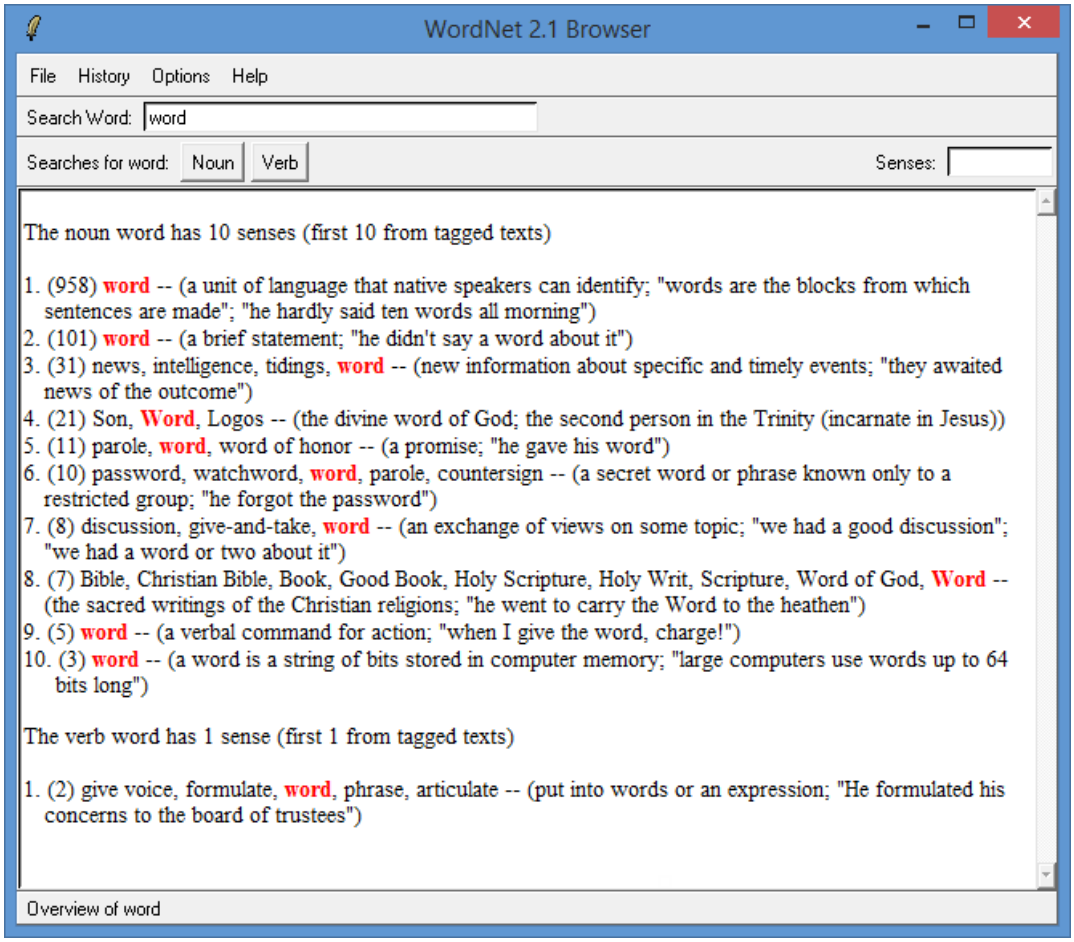

Figura 2: Resultados da pesquisa da palavra *word* na WordNet (retirado de Miller (1995)).

Sendo constituída de hierarquias de generalização e especialização, a WordNet facilita a busca por relações dos conceitos do domínio de interesse. A Figura 3 apresenta as opções oferecidas pela WordNet ao se pesquisar pelo conceito de *vitamin*. Percebe-se que que é possível pesquisar tanto conceitos descendentes de *vitamin* (opções *Hyponyms (...is a kind of vitamin), brief* e *Hyponyms (...is a kind of vitamin), full*) quanto conceitos ascendentes (opção *Hypernyms (...)*). Partindo dos resultados destas pesquisas, a extração de relações de especialização mostra-se bastante simples, visto que a maneira como a WordNet organiza seus conceitos é facilmente traduzida a fim de ser inserida em uma ontologia, dado que esta última também é organizada de maneira hierárquica.

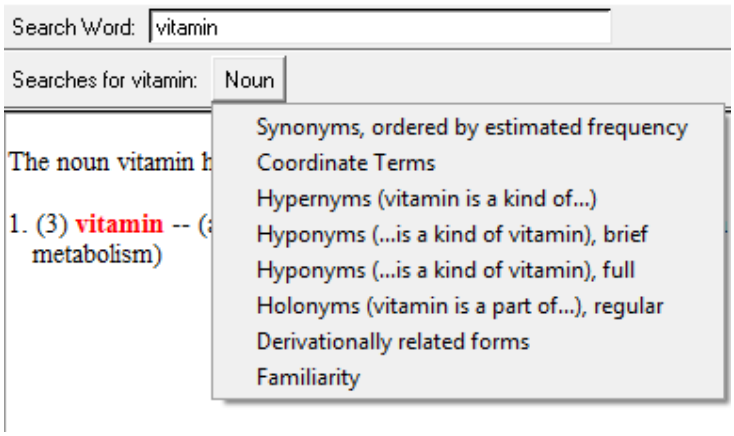

Figura 3: Opções de pesquisa da WordNet (retirado de Miller (1995)).

Outra informação oferecida pela WordNet são os textos explicativos que acompanham os conceitos. Tais textos apresentam breves descrições dos conceitos e sinônimos aos quais se referem, esclarecendo seus significados e trazendo relações de associação com outros conceitos do mesmo domínio ou domínios relacionados. A partir do texto explicativo do conceito de *vitamin A*, que pode ser observado na Figura 1, é possível caracterizar, por exemplo, algumas relações relevantes ao domínio de interesse:

- *Vitamin A is essential for normal vision;*
- *Vitamin A prevents night blindness;*
- *Vitamin A prevents inflammation of the eyes;*
- *Vitamin A prevents dryness of the eyes;*

De maneira geral, os textos explicativos da WordNet expõem as informações de maneira clara e direta, facilitando a extração das relações de associação. Por outro lado, a característica de serem sucintos limita a quantidade de relações de associações

que podem ser extraídas dos textos explicativos. Existe, portanto, a necessidade de fontes de informações mais completas sobre o domínio de interesse. Contudo, a WordNet propicia um meio estruturado para a extração de relações de especialização, que facilita a construção de hierarquias.

2.2. O Conceito de Ontologias

Na pesquisa de Corcho, Fernández-Lopéz e Gómez-Pérez (2003) é apresentada uma grande variedade de definições do conceito de ontologia que existem na literatura. Apesar disso, há um consenso quanto ao significado do conceito de ontologia, de modo que o seu uso não acontece de maneira errônea.

Buitelaar e Magnini (2005) definem ontologias como representações de um domínio de maneira formal, de modo que tal representação segue um padrão estabelecido dentro de uma comunidade e pode ser interpretada por máquinas. Horridge (2011), por sua vez, diz que ontologias são estruturas que representam o conhecimento do domínio de interesse descrevendo os conceitos de tal domínio bem como as relações existentes entre os conceitos.

Não há, na literatura, um acordo sobre como deve ser a representação formal de uma ontologia. Dicionários, *thesauri* ou taxonomias podem ser considerados ontologias, dependendo do ponto de vista (LEHMANN; VÖLKER, 2014).

No entanto, muitas das metodologias para construção de ontologias, apresentadas na seção 1.4, definem que ontologias são organizadas de maneira hierárquica, constituídas de classes e subclasses representadas por relações de especialização. Além disso, ontologias também possuem relações que indicam características ou propriedades dos indivíduos que as possuem (CARVALHEIRA, 2007), definidas como relações de associação.

As relações existentes em uma ontologia são definidas como binárias pois cada uma interliga somente dois conceitos dentro do domínio, sem a influência de um terceiro. Por exemplo, a relação *vitamin a – can be found in – carrot*, presente no domínio de interesse, não depende de um terceiro participante para ser verdadeira ou fazer sentido, de modo que é considerada binária.

Similar à estrutura utilizada pela WordNet, ontologias suportam consultas baseadas na semântica de conceitos e relações de determinado domínio, em vez de consultas baseadas apenas em palavras chaves (DONG; HUSSAIN; CHANG, 2008)

Em uma ontologia baseada em relações de especialização, os conceitos do domínio de interesse são organizados de maneira hierárquica, podendo estes conceitos serem definidos como classes e subclasses. Um conceito de uma ontologia herda todos os atributos e características de seus conceitos ascendentes (UNNI; BASKARAN, 2013). O exemplo abaixo apresenta uma relação simples de especialização dentro do conceito de doença, presente no domínio de interesse:

Beriberi

Wet Beriberi

Dry Beribery

A partir de tal relação de especialização, se considerarmos a existência da relação de associação *beriberi – can be provoked by – vitamin B1 deficiency*, é possível definir, dentro de uma ontologia, que os conceitos *wet beriberi* e *dry beriberi* também estão interligados a *vitamin b1 deficiency* por meio da relação *can be provoked by*.

O exemplo anterior demonstra uma relação de associação que interliga dois conceitos presentes em hierarquias diferentes dentro de uma ontologia. *Beriberi* é uma doença, e portanto possui relação de especialização com o conceito de *disease*. Por outro lado, *vitamin B1 deficiency* é considerada uma causa, e, deste modo, é uma especialização do conceito *cause*. Interligando conceitos de diferentes hierarquias por meio de associações, ontologias são construídas como estruturas que organizam, formalmente, as informações sobre o domínio selecionado.

3. Desenvolvimento do Trabalho Proposto

O processo utilizado para o desenvolvimento deste trabalho foi construído a fim de identificar um método manual de extração e construção de ontologias, com o intuito de analisar o potencial do encadeamento de relações dentro do domínio de interesse no que tange à identificação de relações não explícitas nas fontes originais.

O processo é ilustrado pela Figura 4, que apresenta de maneira encadeada os passos seguidos pelo desenvolvimento deste trabalho A partir da escolha do domínio de interesse, cuja motivação é explicada na seção 1.2, foram escolhidas as fontes de informações sobre as quais serão realizadas a extração de relações. Optou-se pelo foco no conceito de doenças nutricionais, cujas relações envolvem os conceitos de sintomas, doenças, causas, nutrientes e alimentos. As etapas do processo são descritas nas seções que seguem.

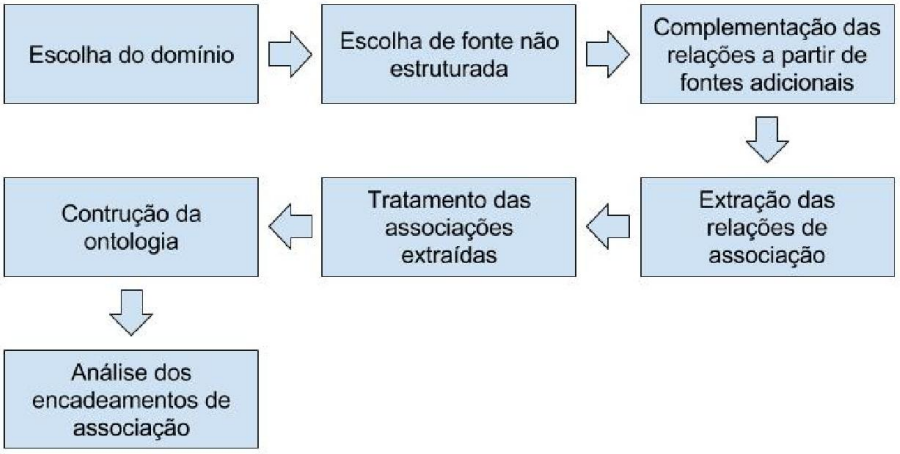

Figura 4: Processo de construção da ontologia para análise de encadeamentos de associação (elaborado pelo autor).

3.1. Escolha da Fonte para Extração de Relações de Associação

Inicialmente foi determinado o domínio do trabalho e os conceitos relevantes dentro do domínio de interesse, que possibilitariam a prototipação do encadeamento de conceitos com o intuito de realizar pesquisas sobre tal domínio. Os conceitos a partir dos quais foram caracterizadas as relações de especialização a serem incluídas na

ontologia são: *symptom – disease – cause – nutriente – food*. De maneira geral, as relações de associação relacionam os conceitos na ordem lógica listada, uma vez que esta é a progressão natural dentro do domínio: um sintoma é característico de uma doença, que se desenvolve devido a alguma causa, a qual pode ser prevenida por determinado nutriente que é encontrado em algum alimento.

Com os conceitos de interesse definidos, pode-se pensar na extração dos relações de associação entre os conceitos selecionados. Os textos explicativos da WordNet contém algumas relações de associação relevantes ao domínio de interesse. Por exemplo, a análise dos textos explicativos sobre Vitamin B12 e Vitamin B2, presentes na Figura 5, permite a identificação das seguintes relações de associação:

- *Vitamin B12 – is used to treat – pernicious anemia*
- *Vitamin B2 – prevents – skin lesion*
- *Vitamin B2 – prevents – weight loss*

=> vitamin B12, cobalamin, cyanocobalamin, antipernicious anemia factor -- (a B vitamin that is used to treat pernicious anemia)
=> vitamin B2, vitamin G, riboflavin, lactoflavin, ovoflavin, hepatoflavin -- (a B vitamin that prevents skin lesions and weight loss)

Figura 5: Textos explicativos da WordNet sobre *Vitamin B12* e *B2* (retirado de Miller (1995)).

As relações encontradas nos textos explicativos da WordNet são claras e podem ser extraídas facilmente pois, de maneira geral, as frases são construídas de forma simples. As relações devem ser extraídas pensando-se nos conceitos os quais interligam. Nos exemplos acima, a primeira associação relaciona um conceito descendente de *nutrient* com um conceito descendente de *disease*. As outras duas, por sua vez, relacionam *nutrient* com *symptom*.

No que diz respeito a relações de associação, a WordNet mostra-se relativamente limitada, pois os textos explicativos contém apenas breves explicações sobre os conceitos aos quais se referem. Para os objetivos deste trabalho, as associações extraídas da WordNet são insuficientes para a prototipação do encadeamento dos conceitos selecionados. Isso se dá pois, para que a ontologia a ser construída possua associações em quantidade o suficiente para possibilitar o encadeamento, é necessário uma fonte mais robusta.

Com o intuito de extrair um conjunto de relações de associação relevantes entre os conceitos selecionados, decidiu-se pelo uso da seção sobre doenças nutricionais da Enciclopédia Britânica (WEININGER, 2015), já citada anteriormente. Essa fonte

relaciona, em tópicos bem definidos, nutrientes e doenças, ocasionalmente trazendo, também, informações sobre sintomas e alimentos como fontes de nutrientes. A Figura 6 apresenta a primeira frase do texto *Nutritional Disease* sobre o nutriente *niacin*, definido pela WordNet como *"a B vitamin essential for the normal function of the nervous system and the gastrointestinal tract"* (MILLER, 1995).

Figura 6: Texto em linguagem natural sobre o nutriente *niacin* (retirado de Weininger (2015)).

Percebe-se que, a partir do texto da enciclopédia, é possível definir relações de associação que não estão presentes no texto explicativo da WordNet, tais como:

- *Niacin deficiency – results in – pellagra*
- *Pellagra – is caused by – niacin deficiency*
- *Niacin – prevents – pellagra*

A escolha de tal fonte se dá, portanto, devido ao seu aspecto concentrado no domínio de interesse. As informações são apresentadas de maneira didática, de modo a serem informativas para, inclusive, não especialistas, e não se distanciam dos conceitos selecionados, de modo que encadeamentos de associações são naturalmente construídos a partir das associações extraídas. Além disso, o encadeamento de associações requer que uma fonte robusta seja utilizada na extração das relações, já que uma ontologia com poucas relações associativas consequentemente apresentaria poucos encadeamentos.

3.2. Complementação com Relações Adicionais

Uma vez que a fonte das relações de associação foi selecionada, deu-se início à seleção das relações de especialização para a construção da ontologia. A WordNet foi escolhida para a extração das relações de especialização, uma vez que sua organização hierárquica é facilmente transposta para uma ontologia, que é estruturada de maneira similar.

A partir da WordNet, obteve-se a hierarquia de especialização do conceito chave, ou raiz, *nutrient.* Uma amostra da hierarquia de *nutrient*, extraída manualmente da WordNet, é apresentada a seguir:

Nutrient

Vitamin
- *Vitamin A*
  - *Vitamin A1*
  - *Vitamin A2*
- *Vitamin D*
- *Vitamin E*
  - *Alpha-tocopheral*
- *Vitamin F*
- *Vitamin K*
  - *Vitamin K1*
  - *Vitamin K2*

As hierarquias de especialização dos outros conceitos raízes (*symptom, disease, cause* e *food*), foram construídas a partir dos conceitos presentes nas relações de associação extraídas da fonte de informação não estruturada da Enciclopédia Britânica. Isto é, os conceitos identificados nas associações (nomes de doenças, sintomas, alimentos, etc) foram organizados hierarquicamente sob o conceito raiz do qual fazem parte. A Figura 7 apresenta um extrato da fonte pesquisada que dispõe sobre o nutriente *thiamin*. Uma vez que as relações de association forem construídas, é possível identificar, por exemplo, que os conceitos de *wet beriberi* e *dry beriberi* devem estar subordinados hierarquicamente ao conceito de *beriberi*, que, por sua vez, é subordinado ao conceito raiz *disease*.

Prolonged deficiency of thiamin (vitamin B_1) results in beriberi, a disease that has been endemic in populations where white rice has been the staple. Thiamin deficiency is still seen in areas where white rice or flour constitutes the bulk of the diet and thiamin lost in milling is not replaced through enrichment. Symptoms of the form known as dry beriberi include loss of appetite, confusion and other mental symptoms, muscle weakness, painful calf muscles, poor coordination, tingling and paralysis. In wet beriberi there is edema and the possibility of an enlarged heart and heart failure. Thiamin deficiency can also occur in populations eating large quantities of raw fish harbouring intestinal microbes that contain the enzyme thiaminase. In the developed world, thiamin deficiency is linked primarily to chronic alcoholism with poor diet, manifesting as Wernicke-Korsakoff syndrome, a condition with rapid eye movements, loss of muscle coordination, mental confusion, and memory loss.

Figura 7: texto da Enciclopédia Britânica sobre o nutriente *thiamin* (retirado de Weininger (2015)).

Uma relação importante para este trabalho, *nutriente – food*, mostrou-se presente apenas esporadicamente na seção sobre doenças nutricionais da Enciclopédia Britânica. A fim de alimentar a ontologia com relações de associação entre esses dois conceitos, utilizou-se de uma fonte estruturada contendo os alimentos mais ricos em cada nutriente. O *site* "http://www.healthaliciousness.com" apresenta, de maneira organizada, os dez alimentos com maior concentração de cada um dos nutrientes de interesse para este trabalho.

Por fim, foi necessário identificar as relações de associação entre *cause* e *nutrient*, que, apensar de intuitivas (a deficiência de uma vitamina acontece devido a falta da mesma), não estavam explicitadas na fonte pesquisada.

3.3. Extração das Relações de Associação

Com a hierarquia inicial definida, deu-se início à extração das relações de associação a partir da fonte selecionada. Para tal, procurou-se manter o sentido das relações apesar das construções sintáxicas complexas presentes em textos de linguagem natural.

As relações de associação foram extraídas manualmente e organizadas conforme apresenta a Figura 8. As relações adicionais, definidas na seção 3.2, foram definidas a seguir, a fim de complementar a ontologia e possibilitar a análise sobre o encadeamento de associações.

Vitamin E deficiency	*can develop in*	premature infants	-	-
		people with impaired fat absorption or metabolism	-	-
	can cause	fragility of red blood cells (hemolysis)	-	-
	can cause	neuromuscular dysfunction involving the spinal cord	*results in*	loss of reflexes
				impaired balance
				impaired coordination
				muscle weakness
		neuromuscular dysfunction involving the retina	*results in*	visual disturbances

Figura 8: relações de associação da causa *Vitamin E deficiency* extraídas da Enciclopédia Britânica (elaborado pelo autor).

Os textos da Enciclopédia Britânica são complexos e apresentam maiores dificuldades na extração de associações quando comparados com os textos explicativos da WordNet. A construção das frases se dá de maneira menos direta, e, apesar da clareza das informações, a extração de relações binárias requer esforço na interpretação do texto para que não ocorra perda de sentido.

3.4. Tratamento das Relações Extraídas

Relações de associação extraídas a partir de fontes não estruturadas precisam passar por um tratamento a fim de padronizá-las, possibilitando sua inclusão em uma ontologia que permita pesquisas sobre o encadeamento das associações. Nas relações de

associação extraídas de textos em linguagem natural, frequentemente um mesmo tipo de relação está escrita de maneira e com palavras diferentes. Um exemplo em que essa situação é bastante comum são as relações entre causas e sintomas:

- *Vitamin A deficiency – can cause – night blindness*
- *Vitamin A deficiency – is the leading cause of – blindness in child*
- *Vitamin B12 deficiency – can result in – nerve degeneration*

Percebe-se, então, a necessidade de agrupar associações com semântica e significados similares em apenas um nome de relação representativo do grupo de associações, sem grande perda de sentido. Sobre o grupo de associações demonstrado acima, por exemplo, foi escolhida a relação *can cause* a fim de padronizar todas as relações entre uma causa e um sintoma. Outros nomes de associação escolhidos para representar relações entre os conceitos incluem:

- *Can manifest as: cause – disease*
- *Can be found in: nutrient – food*
- *Is characterized by: disease – symptom*
- *Can be treated by: disease – nutrient*
- *Is provoked by fault of: cause – nutrient*

Além disso, é necessária a definição de nomes de relações contrários aos definidos com base nos textos das quais as associações foram extraídas. Isso se dá pois o encadeamento dos conceitos deve acontecer nos dois sentidos das associações, de modo que este pode partir tanto de uma extremidade (*symptom*) quanto da outra (*food*). As relações inversas das relações listadas acima são:

- *Can cause – can be caused by*
- *Can manifest as – can be provoked by*
- *Can be found in – contains*
- *Is characterized by – is characteristic of*
- *Can be treated by – can treat*
- *Is provoked by fault of – when in fault can provoke*

Alguns exemplos de conceitos interligados pelas associações descritas acima incluem:

- *Vitamin B6 deficiency – can cause / can be caused by – weakness*
- *Vitamin D deficiency – can manifest as / can be provoked by – rickets*
- *Vitamin A – can be found in / contains – carrot*

- *Scurvy – is characterized by / is characteristic of – joint pain*
- *Pernicious anemia – can be treated by / can treat – vitamin B12*
- *Vitamin C deficiency – is provoked by fault of / when in fault can provoke – vitamin C*

Frequentemente relações de associação estão definidas em textos de linguagem natural por meio de construções sintáticas complexas, que dificultam na extração de relações de associação binárias. Essa dificuldade acarreta na eventual perda de pelo menos parte das informações originais contidas nas frases das fontes de consulta. A Figura 9 exibe um exemplo de associação complexa. A partir da frase destacada, pode-se definir a associação: *injection of vitamin K within six hours of birth – can protect against – hemorrhagic disease of the newborn*. Contudo, não há como representar em uma simples relação binária a condição "*within six hours of birth*", que qualifica a associação e é necessária para que a ela seja totalmente verdadeira, sem perda de informação. Frases complexas como esta são frequentes na literatura do domínio de interesse.

VITAMIN K

Vitamin K is necessary for the formation of prothrombin and other blood-clotting factors in the liver, and it also plays a role in bone metabolism. A form of the vitamin is produced by bacteria in the colon and can be utilized to some degree. Vitamin K deficiency causes impaired clotting of the blood and internal bleeding, even without injury. Due to poor transport of vitamin K across the placenta, newborn infants in developed countries are routinely given the vitamin intramuscularly or orally within six hours of birth to protect against a condition known as hemorrhagic disease of the newborn. Vitamin K deficiency is rare in adults, except in syndromes with poor fat absorption, in liver disease, or during treatment with certain anticoagulant drugs, which interfere with vitamin K metabolism. Bleeding due to vitamin K deficiency may be seen in patients whose gut bacteria have been killed by antibiotics.

Figura 9: Texto em linguagem natural sobre o nutriente *Vitamin K* (retirado de Weininger (2015)).

O agrupamento de associações com sentido semelhante em um único nome de associação, explicado anteriormente, também deve ser feito de maneira cuidadosa. Ainda que o intuito não seja o de alterar o sentido da relação, a mudança na semântica pode resultar em perda de informação, já que fatores como intensidade (*may cause*, *can cause* e *cause*, por exemplo) são relevantes, especialmente no domínio de interesse deste trabalho, a nutrição.

3.5. Construção da Ontologia

Finalizadas as extrações das relações de especialização e associação, foi construída uma ontologia a ser utilizada no estudo de caso sobre os encadeamentos entre os conceitos. A ontologia possui hierarquias dos cinco conceitos principais selecionados do domínio (*symptom, disease, cause, nutriente* e *food*), e é composta pelas relações de associação extraídas a partir dos textos em linguagem natural da fonte especificada na seção 3.1, tratadas conforme descrito na seção 3.3.

Uma vez que todas as relações relevantes ao domínio estão elencadas, a construção de uma ontologia é uma tarefa trivial, bastando que seja utilizada uma representação e ferramenta de escolha para tal. Ferramentas como Protégé (disponível em "http://protege.stanford.edu/"), por exemplo, disponibilizam um ambiente para construção de ontologias com uma grande gama de funcionalidades. Este trabalho utiliza uma representação e ferramenta própria, desenvolvidas na pesquisa de Tulio e Batista (2015), a fim de construir a ontologia do domínio de interesse. A interface da ferramenta é apresentada na Figura 10.

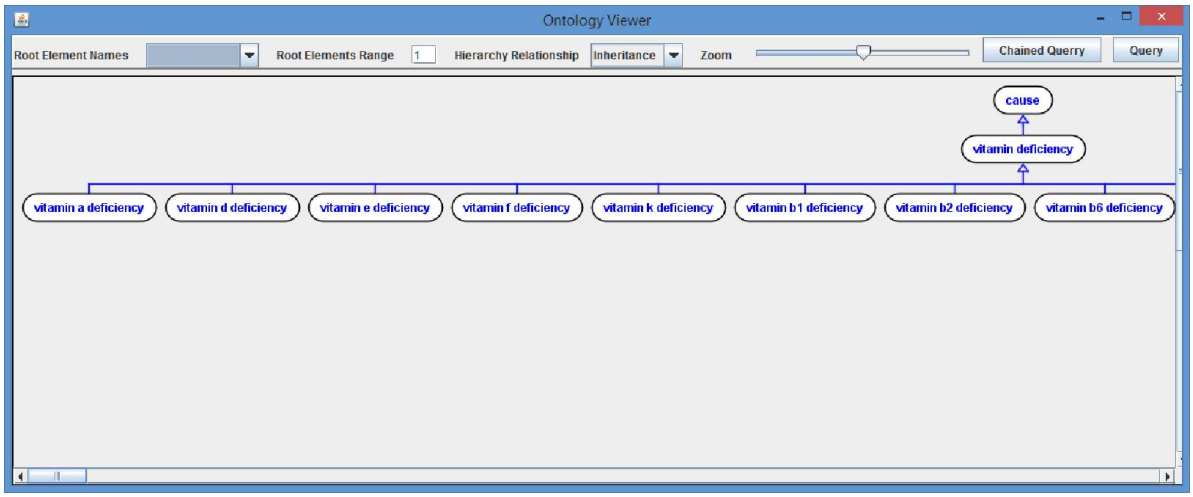

Figura 10: Interface gráfica da ferramenta de visualização de ontologias (elaborado pelo autor).

3.5.1. Representação Própria e Visualização Gráfica da Ontologia

A representação de ontologias de Tulio e Batista (2015) foi selecionada pois atende as necessidades deste trabalho quanto à representação de uma ontologia, neste caso composta apenas de conceitos relacionados por associação ou especialização. Essa

representação organiza hierarquicamente os conceitos da ontologia, armazenando em cada conceito as relações de associações das quais o conceito participa. Assim como a extração das relações, a construção da ontologia se dá manualmente, sendo acrescentadas, inicialmente, todas as hierarquias de conceitos e, a seguir, as relações de associação.

O trabalho desenvolvido em Tulio e Batista (2015) produziu também a ferramenta de visualização gráfica de ontologias que funciona com a representação própria. O Visualizador de Ontologia foi utilizado pelo presente trabalho para visualizar de maneira clara a ontologia construída, facilitando a busca de conceitos dentro das hierarquias da ontologia. Além disso, a ferramenta gráfica possui funcionalidades quanto à visualização e pesquisa dos encadeamentos de associação, desenvolvidas em Tulio e Batista (2016) e explicadas na seção 3.6.

A Figura 11 ilustra a representação gráfica da hierarquia do conceito de *nutrient*, conforme exibida pelo Visualizador de Ontologia. A Figura 12 mostra, também graficamente e utilizando-se da capacidade do visualizador de exibir associações, as relações de *vitamin C* com conceitos da hierarquia de alimentos.

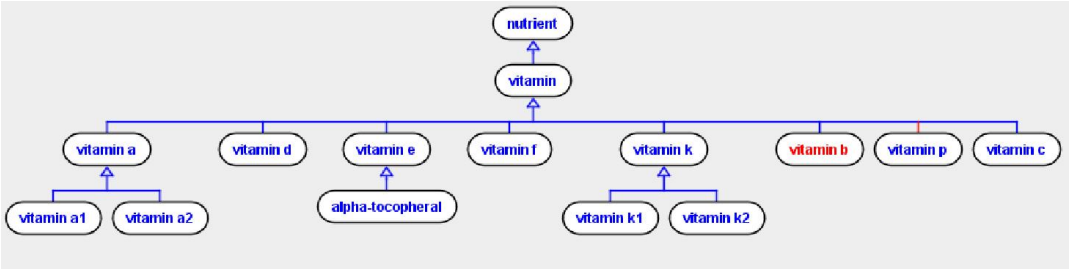

Figura 11: Representação gráfica da hierarquia de *nutriente* (elaborado pelo autor).

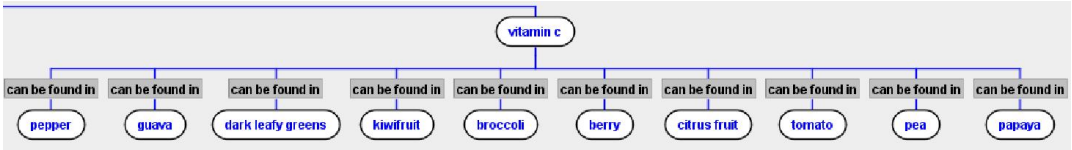

Figura 12: Relações de associação que interligam *Vitamin C* com alimentos (elaborado pelo autor).

3.6. Encadeamentos de Associações representados na Ontologia

Uma vez que a ontologia encontra-se completa, pode-se perceber como as relações de associação extraídas da fonte em linguagem natural formam, em conjunto com as associações incluídas a fim de complementar a ontologia, encadeamentos como:

- *Weak muscles – is characteristic of – osteomalacia – can be provoked by – vitamin D deficiency – is provoked by fault of – vitamin D – can be found in - egg*

Esse encadeamento de associações passa por todas as hierarquias definidas: *symptom, disease, cause, nutrient* e *food*. A partir dele é possível verificar que o sintoma *weak muscles* pode ser prevenido pelo alimento *egg*, ou ainda que a doença *osteomalacia* é combatida pelo nutriente *vitamin D*.

Encadeamentos de associações menores também podem ser encontrados, por exemplo:

- *Dry skin – can be caused by – vitamin A deficiency – is provoked by fault of – vitamin A – can be found in – carrot*

Nesse caso, o encadeamento de associações não passa pelo conceito de *disease*, partindo de sintoma (*dry skin)* diretamente para causa (*vitamin A deficiency*). Encadeamentos como esses permitem responder perguntas dentro do domínio de interesse que, quando dentro de um texto em linguagem natural nem sempre estão facilmente acessíveis.

3.6.1. Ferramenta de Visualização de Encadeamentos de Conceitos

A fim de visualizar os encadeamentos de associação da ontologia, foi utilizada a funcionalidade de Visualização de Encadeamentos de Conceitos presente na ferramenta referenciada anteriormente. Tal funcionalidade foi desenvolvida na pesquisa de Tulio e Batista (2016). A Figura 13 apresenta a interface gráfica da funcionalidade.

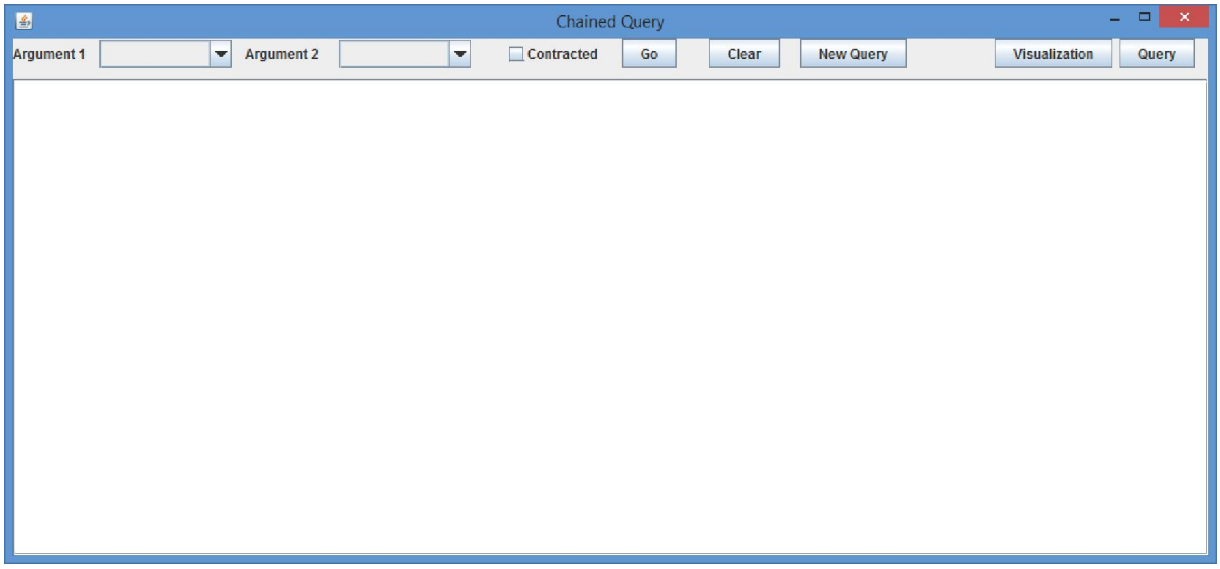

Figura 13: Interface gráfica da funcionalidade de visualização de encadeamento de conceitos (elaborado pelo autor).

A ferramenta permite que sejam encontrados, automaticamente, encadeamentos de associações a partir de qualquer conceito da ontologia, esteja este no topo da hierarquia ou não. É possível definir a busca de encadeamentos específicos, ao selecionar os pontos de partida e chegada do encadeamento de interesse, e a busca de encadeamentos em geral, quando apenas o ponto de partida é selecionado, ou são selecionados conceitos abrangentes, ou raízes (no caso deste trabalho, os conceitos *symptom, disease, cause, nutrient* e *food*).

O uso do Visualizador de Encadeamentos de Conceitos por este trabalho auxilia na validação dos encadeamentos da ontologia. A automação da pesquisa sobre os encadeamentos facilita a identificação de informações que antes, devido a conceitos não diretamente relacionados, encontravam-se implícitas.

3.6.2. Análise dos Encadeamentos de Associações

A fim de exemplificar os usos e a relevância do encadeamento de associações, consideremos a seguinte pesquisa realizada sobre a ontologia deste trabalho: quais alimentos previnem o sintoma *confusion*? Utilizando-se da representação da ferramenta de Visualização de Encadeamentos de Conceitos, o resultado de tal pesquisa pode ser observado na Figura 14. Percebe-se que, considerando as relações utilizadas para construir a ontologia deste estudo de caso, o sintoma *confusion* pode ser prevenido por alimentos que contém os nutrientes *Vitamin B6* ou *Vitamin B1*.

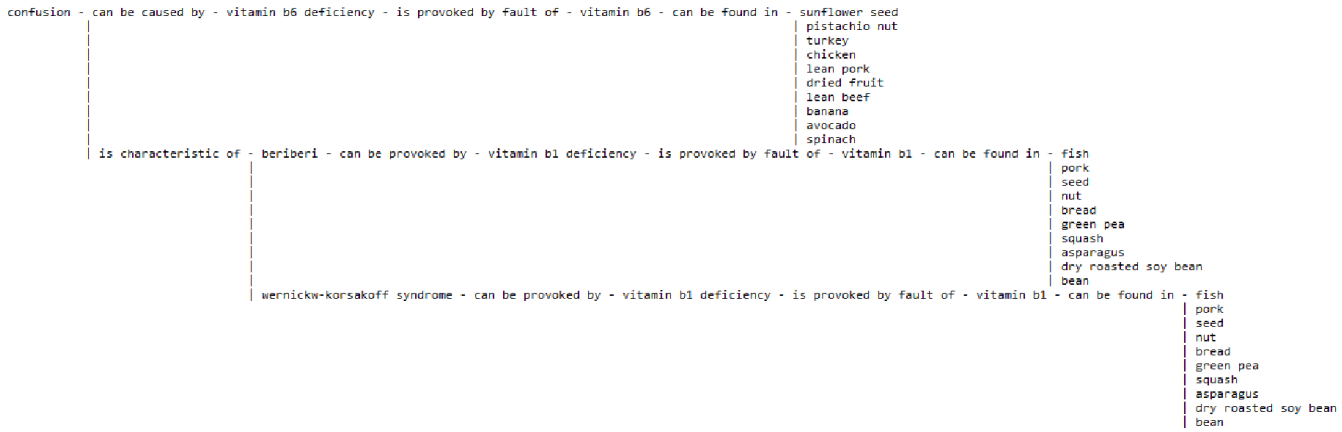

Figura 14: Encadeamento de associações que interligam *confusion* a *food* (elaborado pelo autor).

Os encadeamentos não tem como finalidade garantir que apenas os alimentos elencados na Figura 14 podem ser utilizados para prevenir o sintoma confusão. Isso só seria possível caso a ontologia sobre a qual a pesquisa de encadeamentos é realizada contivesse absolutamente todos os alimentos que servem de resposta para a pesquisa, inseridos hierarquicamente sob o conceito de *food* e relacionados aos outros conceitos pelas relações de associação. Percebe-se que a resposta às pesquisas realizadas sobre os encadeamentos de associações são completas na mesma medida que as fontes de informações são. Verifica-se aqui, portanto, a necessidade de extrair relações a partir de fontes robustas e completas, conforme discutido anteriormente.

O exemplo anterior demonstra encadeamentos que iniciam em uma das extremidades dos conceitos selecionados do domínio e chegam até a outra (de *symptom* até *food*). Isso se dá pois a pergunta original foi feita de maneira que o encadeamento deveria buscar as relações entre esses dois conceitos. Outras pesquisas que retornam encadeamentos menores podem ser realizadas, por exemplo: quais os sintomas da doença *scurvy*? Que alimentos previnem a deficiência de vitamina A1? Quais as doenças prevenidas ou combatidas pelo consumo de frutas cítricas?

Pesquisas como “quais alimentos previnem ou combatem a doença *scurvy*?” requerem que relações de associações inversas tenham sido incluídas na ontologia, conforme descrito na seção 3.4. A Figura 15 apresenta o resultado dessa pesquisa.

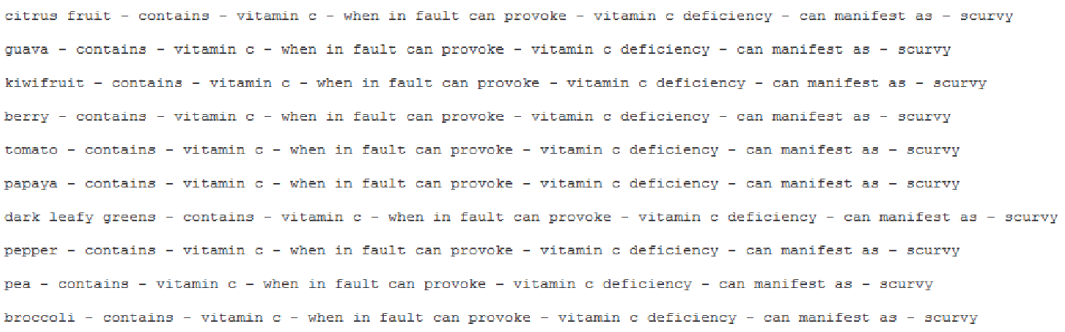

Figura 15: Encadeamentos de associações que ligam *food* com *scurvy* (elaborado pelo autor).

As informações retornadas pelas pesquisas sobre os encadeamentos de associações dificilmente estão claramente definidas nas fontes não estruturadas das quais as relações são extraídas. Informações que envolvem muitos conceitos para serem definidas, conforme os exemplos apresentados nesta seção, dificultam a sua identificação nos textos originais. Os encadeamentos de associação se mostram relevantes nesse sentido pois permitem identificar, automaticamente, relações antes implícitas.

No entanto, para tal fim, é importante que a base de informação estruturada sobre a qual existem os encadeamentos de associação, a ontologia, tenha sido construída utilizando-se de informações robustas e fidedignas. Isto é, a extração das relações a partir das fontes de consulta do domínio de interesse precisa ser feita cuidadosamente, para que informações falsas não sejam incluídas na ontologia e comprometam os resultados.

Neste estudo de caso são apresentados resultados positivos no que tange à corretude das informações identificadas pela prototipação das pesquisas de encadeamentos de associação. Os resultados obtidos pelas pesquisas se mostram corretos quando analisados em relação às informações selecionadas e incluídas na ontologia a partir da fonte não estruturada e das fontes com relações complementares.

A ontologia construída e utilizada por este estudo apresenta apenas uma fonte principal de relações de associação, e apresenta resultados coerentes com as informações as quais possui. Uma ontologia construída a partir de uma quantidade maior de fontes, com relações de associação devidamente tratadas, potencializaria a consulta de relações entre conceitos distantes e, consequentemente, a identificação de informações não explícitas.

4. Considerações Finais

Nesta seção são apresentadas: as conclusões, as dificuldades encontradas e sugestões para trabalhos futuros.

4.1. Conclusões

O encadeamento de associações em uma ontologia bem estruturada tem a função de representar informações de determinado domínio que antes se encontravam em texto de linguagem natural. Para que se extraia conhecimento de tais fontes, é essencial a boa interpretação de texto, bem como a capacidade de correlacionar conceitos e interligá-los a fim de chegar a conclusões. A extração de relações binárias que podem ser incluídas em uma ontologia a partir dos textos de linguagem natural permitem que as informações sejam representadas em uma forma estruturada e de fácil manipulação e validação. Isto é, representar as informações em uma ontologia facilita o acesso ao conhecimento.

Ontologias suportam a realização de pesquisas sobre os domínios de interesse, oferecendo facilidade e suporte a vários tipos de consulta. É possível, por exemplo, pesquisar quais conceitos estão interligados por determinada associação, ou ainda quais são as associações que relacionam determinado conceito aos descendentes de outro elemento da ontologia. Essa flexibilidade possibilita que se chegue até a informação de maneira mais rápida do que seria possível na leitura de um texto em linguagem natural.

O encadeamento de associações traz, ainda, a possibilidade de realizar consultas sobre o domínio que, para serem respondidas, precisam considerar diversos conceitos. A automatização desse processo resulta numa ferramenta que, quando atua sobre uma ontologia bem construída e com relações binárias extraídas adequadamente, permite identificar conhecimentos implícitos, formados por conceitos que não estão ligados diretamente.

Existem, contudo, alguns pontos ainda em aberto que são relevantes para o avanço da pesquisa. O agrupamento de nomes das associações, descrito na seção 3.4, pode ocasionar na perda de informações dos textos originais, comprometendo a integridade das ontologias. Além disso, frases complexas em textos de linguagem natural resultam em associações também complexas que dificultam sua inclusão em uma ontologia.

4.2. Dificuldades Encontradas

Textos em linguagem natural são não estruturados, e requerem interpretação para que o sentido e as relações entre os conceitos sejam identificadas. Essa não objetividade dos textos dificulta a extração de relações de associação binárias sem que haja perda de informação na transição. Relações de associação bem definidas são essenciais para que os encadeamentos de associação funcionem e sejam verdadeiros. Extrair as relações dos textos de linguagem natural é, portanto, a tarefa mais importante e também a mais complexa.

Contudo, a dificuldade durante a identificação de relações de associação no domínio de interesse não se dá somente devido à complexidade característica dos textos em linguagem natural. A não familiaridade com os domínios de interesse também pode limitar a qualidade das informações extraídas.

Além disso, apesar da capacidade das pesquisas sobre os encadeamentos de associações de identificar informações não explicitas, os resultados só serão encontrados caso a ontologia sobre a qual a ferramenta atua esteja completa. Isto é, relações óbvias para o ser humano (no caso do domínio de nutrição: a deficiência de vitamina A é causada pela falta de vitamina A) devem ser incluídas na ontologia, para que o encadeamento de associações não seja "quebrado".

4.3. Trabalhos Futuros

Como trabalho futuro, a extração automática de relações de associação a partir de fontes não estruturadas é uma tarefa desafiadora, que poderá ser utilizada para automatizar a construção de ontologias sobre os domínios de interesse.

A construção de estruturas que suportem as qualificações dos conceitos do domínio, sejam estas de tempo, intensidade ou frequência, viabilizaria a extração de associações mais robustas e sem perda de informações em relação aos textos originais em linguagem natural. Algumas construções semânticas não podem ser transformadas em relações binárias sem que o sentido da relação seja comprometido. A possibilidade de inserir relações complexas em uma ontologia mantendo seu sentido é um passo importante para aperfeiçoar a representação de conceitos e relações em uma ontologia.

REFERÊNCIAS

BERNERS-LEE, T; HENDLER, J; LASSILA, O. **The semantic web.** Scientific American, 2001.

BUITELAAR, P; MAGNINI, B. **Ontology learning from text: An overview**. Ontology learning from text: Methods, evaluation and applications, v. 123 of Frontiers in Artificial Intelligence and Applications. IOS Press, 2005.

CARVALHEIRA, L. C. C. **Método semi-automático de construção de ontologias parciais de domínio com base em textos**. Dissertação de Mestrado, Escola Politécnica de São Paulo, 2007.

CORCHO, O; FERNÁNDEZ-LÓPEZ, M; GÓMEZ-PÉREZ, A. **Methodologies, tools and languages for building ontologies. Where is their meeting point?** Data & Knowledge Engineering 46, 2003.

DONG, H; HUSSAIN, F. K; CHANG, E. **A Survey in Semantic Search Technologies.** In: Second IEEE International Conference On Digital Ecosystems and Technologies, 2008, p. 403-408.

FERNÁNDEZ, M; GÓMEZ-PÉREZ, A; JURISTO, N. **Methontology: From Ontological Art Towards Ontological Engineering**. AAAI Technical Report SS-97-06, 1997.

IQBAL, R; MURAD, M. A. A; MUSTAPHA, A; SHAREF, N. M. **An analysis of ontology engineering methodologies: a literature review.** Research Journal of Applied Sciences, Engineering and Technology. Maxwell Scientific Organization, 2013.

KAPOOR, B; SHARMA, S. **A comparative study ontology building tools for semantic web applications.** International journal of Web & Semantic Technology (IJWesT). Vol. 1, No. 3, 2010.

LEHMANN, J; VÖLKER, J. An introduction to ontology learning. In: LEHMANN, J; VÖLKER, J. **Perspectives on Ontology Learning.** IOS Press, 2014.

MILLER, G. A. **Nouns in WordNet: A Lexical Inheritance System**. Princeton University, 1993.

MILLER, G. A., et al. **Introduction to WordNet: an on-line lexical database**. Princeton University, 1993.

MILLER, G. A. **WordNet: A Lexical Database for English.** Communications of the ACM Vol. 38, No 11: 39-41, 1995

NOY, N; MCGUINESS, D. **Ontology Development 101: A Guide to Creating Your First Ontology.** Stanford University, California, 2001.

ÖHRGEN, A; SANDKUHL, K. **Towards a methodology for Ontology development in small and medium-sized enterprises.** IADIS International Conference on Applied Computing, 2005.

PRINCETON UNIVERSITY. **About WordNet**. WordNet. Princeton University, 2010. Disponível em: <http://wordnet.princeton.edu>. Acesso em: 04 de fev. 2015.

TULIO, L. S; BATISTA, J. J. **Assistente para validação incremental na linguagem OWL de ontologia construída automaticamente**. ENEPEX, 2015.

TULIO, L. S; BATISTA, J. J. **Prototipagem de construção de ontologia para suportar consultas em um subdomínio de nutrição.** ENEPEX, 2016.

UNNI, M; BASKARAN, K. **Designing ontology schema and data instance for nutrition domain.** In: International Journal of Computer Science Issues. Vol. 10, Issue 1, No 2, 2013.

USCHOLD, M; KING, M. **Towards a methodology for building ontologies.** In: Workshop on Basic Ontological Issues in Knowledge Sharing. International Joint Conference on Artificial Intelligence, 1995

WHITBREAD, D. **Top 10 lists for the most nutritious foods**. Disponível em: <https://www.healthaliciousness.com/most-nutritious-foods-lists.php>. Acesso em: 15 jul. de 2015.

WEININGER, J. **Nutritional disease.** In: Encyclopedia Britannica. 2016. Disponível em: <http://global.britannica.com/science/nutritional-disease>. Acesso em: 9 fev. 2015.