XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 – Informação e Tecnologia

SEMANTIZAR: UM SISTEMA DE EXTRAÇÃO DE RELAÇÕES SEMÂNTICAS

SEMANTIZAR: *SEMANTIC RELATIONS EXTRACTION SYSTEM*

Lucinéia Souza - Universidade Federal Ouro Preto;
Gercina Ângela de Lima - Universidade Federal de Minas Gerais

Modalidade: Trabalho Completo

Resumo: As relações semânticas são fundamentais para a compreensão da natureza da ligação entre conceitos em um domínio, tornando as representações do conhecimento mais próximas da realidade. Neste sentido, o objetivo deste artigo é apresentar o sistema web chamado Semantizar. Esse sistema foi desenvolvido para apoiar a extração de relações semânticas para a representação do conhecimento. Para tal, ele utiliza como entrada bases conceituais, elaboradas previamente a partir de documentos acadêmicos e, os respectivos documentos acadêmicos. Como saída, o Semantizar indica a frase onde ocorre os pares de conceitos, cabendo ao usuário validar a relação semântica entre esses conceitos. Um estudo de caso foi realizado para verificar a eficiência do Semantizar no suporte à extração de relações semânticas, os resultados apontam um enriquecimento semântico da base conceitual utilizando o Semantizar pois, na representação decorrente, todos os conceitos da amostra relacionam-se com os demais. Como contribuição, acredita-se que o Semantizar possa atuar como embrionário para a extração automática de relações semânticas, provendo um conjunto de dados de relações semânticas para o aprendizado de máquina, desse modo, ele contribui com pesquisas que envolvem o processamento de linguagem natural.

Palavras-Chave: Relações Semânticas. Extração de Relações Semânticas. Explicitação de Relações Semânticas. Representação do Conhecimento.

Abstract: Semantic relations are fundamental for understanding the nature of connection between concepts in a domain, making representations of knowledge closer to reality. In this sense, the purpose of this paper is to present the web system called Semantizar. This system was developed to support the extraction of semantic relations for knowledge representation. For this, it uses as input conceptual bases, elaborated previously from academic documents and, the respective academic documents. As an output, the Semantizar indicates the sentence where the pairs of concepts occur, and the user validates the semantic relation between these concepts. A case study was carried out to verify the efficiency of Semantizar in the support to the extraction of semantic relations, the results point to a semantic enrichment of the conceptual basis using the Semantizar, because in the resulting representation all the concepts of the sample are related to the others. As a contribution, it is believed that Semantizar can act as an embryo for the automatic extraction of semantic relations, providing a dataset of semantic relations for the machine learning, in this way, it contributes with research that involves natural language processing.

Keywords: Semantic Relations. Extraction of Semantic Relations. Explanation of Semantic Relations. Knowledge Representation.

1 INTRODUÇÃO

As relações semânticas entre os conceitos em um domínio possibilitam criar estruturas de conhecimento organizadas de maneira compreensível e assim facilitar para o usuário assimilar o propósito da associação entre os conceitos no contexto que lhe é apresentado.

No âmbito deste artigo, a representação do conhecimento é traduzida em instrumentos tais como os Sistemas de Organização do Conhecimento (SOC) que têm como base conceitos de um domínio. Em SOC, tais como taxonomias, o tipo de relações semânticas hierárquicas é visivelmente constatado. Em outros, tais como os tesauros, nota-se, além do tipo hierárquico, as relações semânticas associativas e de equivalência.

Identificar os tipos de relações semânticas em um SOC é importante para compreender a posição de um conceito em um domínio específico. Porém, somente a identificação dos tipos de relações semânticas algumas vezes não é suficiente para indicar ao usuário qual é, de fato, a natureza da relação entre os conceitos.

Explicitar as relações semânticas de um SOC manualmente pode ser complexo dependendo do domínio que se pretende representar. Nesse sentido, pesquisas sobre extração automática de relações semânticas podem apoiar a explicitação dessas relações. Porém, essas pesquisas atendem a idiomas específicos e, como cada idioma tem suas características, a adaptação dessas pesquisas de um idioma para outro pode não ser adequada. Neste sentido, este artigo apresenta um sistema Web para a extração de relações semânticas para a representação do conhecimento no contexto do idioma português brasileiro denominado SEMANTIZAR.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta a metodologia de pesquisa apontando as principais características da mesma e os procedimentos utilizados. Na Seção 3 são abordadas as técnicas de extração de relações. A Seção 4 enumera os trabalhos correlatos sobre extração de relações. A Seção 5 apresenta o Semantizar, explicando seu contexto, modelagem de dados e implementação. Na Seção 6 é apresentada a avaliação do Semantizar e são apontadas as suas contribuições. Por fim, a Seção 7 aponta as considerações finais.

2 METODOLOGIA

Quanto à sua natureza, esta pesquisa está classificada como aplicada. Segundo Silveira e Córdova (2009, p. 35), a pesquisa aplicada "[o]bjetiva gerar conhecimentos para a aplicação prática, dirigidos à solução de problemas específicos". No caso deste artigo, essa característica se aplica porque um sistema web de extração de relações semânticas foi criado e implementado.

Quanto aos objetivos da pesquisa, este artigo tem a característica de ser uma pesquisa exploratória, que permitiu o conhecimento dos assuntos que intercederam a proposta da pesquisa.

Quanto aos procedimentos, definiu-se a realização da pesquisa bibliográfica e do estudo de caso. A pesquisa bibliográfica, como o próprio nome sugere, busca em fontes diversas, tais como livros e artigos científicos, suportes para fundamentação teórica (PRADONOV; FREITAS, 2013). Neste artigo, a pesquisa bibliográfica foi realizada em dois momentos: (1) no levantamento bibliográfico da fundamentação teórica-metodológica e; (2) no levantamento de trabalhos correlatos sobre extração de relações.

Com relação ao estudo de caso, este foi utilizado para avaliar a eficiência do Semantizar. A metodologia para a realização do estudo de caso foi organizada de acordo com a proposta de Processo de Experimentação de Wohlin et al. (2000), porém adaptado para o artigo.

Quanto ao universo de pesquisa, utilizou-se na experimentação estruturas classificatórias e seus respectivos documentos acadêmicos do tipo tese ou dissertação. Nesse universo, a amostra empregada foi a estrutura facetada, o Mapa hipertextual MHTX de Lima (2004) e a respectiva tese por ele representada: *Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores*, de Naves (2000).

O MHTX é um modelo de navegação hipertextual em contexto criado por Lima (2004) para organizar Teses e Dissertações, visando apoiar a leitura e a recuperação desses documentos em Bibliotecas Digitais de Teses e Dissertações. No protótipo do MHTX, Lima (2004) criou três ferramentas de navegação: o sumário expandido, o mapa conceitual e a estrutura facetada. Na implementação, ela utilizou a tese supracitada de Naves (2000) para instanciar os instrumentos criados. Desses instrumentos, a estrutura facetada apresenta as características da amostra desejada para o Semantizar.

Para tornar mais eficiente a aplicabilidade da amostra, um recorte foi realizado na estrutura facetada, uma vez que a quantidade de conceitos não é o principal fator nesse momento. Dessa forma, os assuntos da faceta Personalidade serão utilizados no estudo de

caso. Da mesma maneira, decidiu-se por um recorte da tese relativa à estrutura facetada utilizada. Logo, optou-se por considerar os capítulos 2, 3 e 4 da tese de Naves (2000). Essa escolha decorreu do fato de esses capítulos referirem-se à parte de definições conceituais da tese em questão, por isso eles parecem ser mais relevantes para o propósito ao qual serão aplicados.

3 EXTRAÇÃO DE RELAÇÕES

A Extração de Relações (ER) é uma das temáticas da Extração de Informação (EI), que por sua vez é uma tarefa do Processamento de Linguagem Natural (PLN). O PLN é o estudo científico de linguagens naturais na perspectiva computacional (KUMAR, 2011).

As pesquisas que envolvem o PLN trabalham com dois tipos de sistemas: os que convertem informações de bases de dados computacionais em uma linguagem legível para humanos, que são os sistemas de geração de linguagem natural; e os sistemas de entendimento de linguagem natural, que convertem exemplos da linguagem humana em representações formais para manipulação de programas de computador (KUMAR, 2011). Nesse último caso, uma das tecnologias de PLN é a EI.

A EI extrai automaticamente informações estruturadas, tais como entidades, atributos que descrevem entidades e relacionamentos entre entidades, a partir de fontes não estruturadas. As entidades são tipicamente sintagmas nominais, que denotam nomes de pessoas, locais, organizações, etc. (SARAWAGI, 2008). Nesse caso, no que se refere à identificação de entidades em textos, a EI conta com a tarefa de Reconhecimento de Entidade Nomeada (REN). O REN identifica os nomes de pessoas, organizações, tempo, moeda e expressões de porcentagem em textos por meio de identificadores (GRISHMAN & SUNDHEIM, 1996).

Uma vez que as entidades são reconhecidas no REN, é importante estabelecer a relação entre elas. Nesse caso, alguns dos recursos que podem ser utilizados são: *Surface Tokens* (Tokens de Superfície), *Part Of Speech tags* (POS *tags*) (Marcação de Classe Gramatical), *Syntactic Parse Tree Structure* (Estrutura de Árvore de Análise Sintática) e, *Dependency Graph* (Grafo de Dependência) (SARAWAGI, 2008).

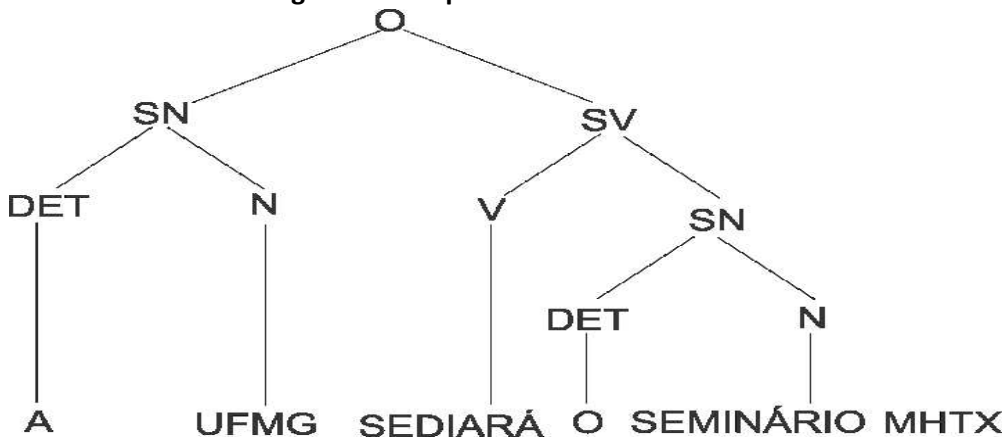
Os Tokens de Superfície são *tokens* ao redor e entre duas entidades que podem sugerir relacionamentos entre elas (SARAWAGI, 2008). Por exemplo, dada a sentença: “UFMG está

localizada em Belo Horizonte.” Ao detectar que UFMG é uma instância para a entidade ORGANIZAÇÃO e Belo Horizonte é uma instância da entidade LOCAL, entre elas existe o *token* de trigrama “está localizada em” que é uma sugestão da existência de um relacionamento entre as entidades ORGANIZAÇÃO e LOCAL, dado que, nesse trigrama existe a presença do radical “localiz” para o verbo localizar, que é um dos verbos recomendados para estas entidades.

As POS *tags* são marcações que denotam a classe gramatical das palavras em uma sentença. Portanto, elas indicam se determinada palavra é um verbo, pronome, adjetivo, advérbio, etc. Nesse caso, quando existe a indicação de um verbo, há um forte indício de uma relação, ou seja a presença de um verbo em uma frase é fundamental para definir uma relação entre entidades (SARAWAGI, 2008). Por exemplo, na sentença: “A UFMG sediará o Seminário MHTX.” Nesse caso, utilizando POS *tags*, a sentença seria marcada da seguinte forma: “A/AD UFMG/SP sediará/VB o/AD Seminário MHTX/SC.” Onde AD é a sigla para Artigo Definido; SP – Substantivo Próprio; VB – Verbo e; SC – Substantivo Composto. Como a palavra “sediará” está marcada como verbo e está entre dois substantivos (UFMG e Seminário MHTX), pode-se afirmar que sediará é a relação entre UFMG e Seminário MHTX.

Enquanto as POS *tags* determinam a classe gramatical das palavras (análise morfológica), a *Syntactic Parse Tree Structure* determina a função sintática das palavras de acordo com seus constituintes, ou seja, de acordo com seu papel dentro da oração, de um ponto de vista estrutural, isto é, de sujeito, objeto, predicado verbal, predicado nominal, adjunto adverbial, etc. A Figura 1 representa uma árvore de análise sintática da oração “A UFMG sediará o Seminário MHTX”, em que O indica oração, SN – sintagma nominal, SV – sintagma verbal, DET – determinante, N – nome e V – verbo. Da mesma forma como na utilização das POS *tags* o verbo indica o relacionamento.

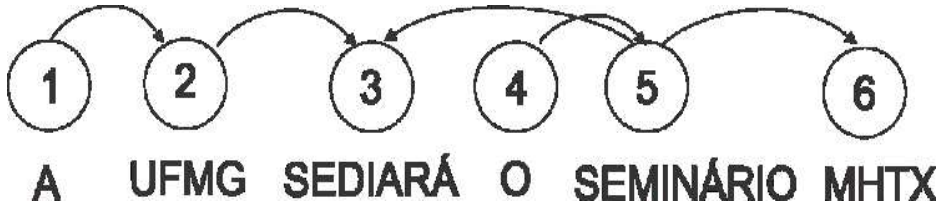
Figura 1: Exemplo de árvore de análise sintática.



Fonte: Extraído de Maia (2018).

Por fim, apresenta-se o *Dependecy Graph*, que, de acordo com Sarawagi (2008), é tipo de grafo direcionado que representa as dependências dos objetos uns com os outros. Nesse sentido, um grafo de dependência liga cada palavra às palavras que dependem dela. Ele é frequentemente utilizado por ser tão adequado quanto uma árvore de análise sintática. No exemplo da Figura 2, cada nó representa uma palavra e as arestas representam a dependência entre as palavras. A palavra "sediará" é ligada e depende de ambos, do nó que representa UFMG e Seminário, o que pode indicar a relação entre UFMG e Seminário.

Figura 2: Exemplo de grafo de dependência.



Fonte: Extraído Maia (2018).

Todos os recursos apresentados nesta seção podem auxiliar na indicação de relações semânticas, pois, de acordo com Sarawagi (2008), eles fornecem "pistas" da existência de um relacionamento em um fragmento de texto.

4 TRABALHOS CORRELATOS

Esta seção apresenta parte do resultado de uma revisão de literatura realizada para apurar o estado da arte das pesquisas sobre extração de relações semânticas. Para tal, utilizou-se a metodologia de revisão de literatura de Okoli e Schabram (2010). Essa metodologia possui oito etapas: (1) proposta de revisão de literatura; (2) protocolo e treinamento; (3) busca na literatura; (4) tela prática; (5) avaliação de qualidade; (6) extração de dados; (7) análise dos achados; (8) escrita da revisão.

Nas etapas iniciais da revisão de literatura, determinou-se, entre outras coisas, a temporalidade da pesquisa, limitada a artigos publicados entre o período de 2013 a 2017, e as bases de dados: *Library, Information Science & Technology Abstracts with Full Text* (EBSCO), *Information Science & Technology Abstracts* - ISTA (EBSCO), *Library and Information Science Abstracts* - LISA (ProQuest), *Web Of Science* e *Scopus,* cujas buscas resultaram em 9.386 artigos. Esses artigos passaram por avaliações e análises de relevância para a pesquisa. Dessa

forma, após as etapas 3, 4, 5 e 6, dos artigos recuperados selecionou-se nove artigos. Desses nove artigos, seis foram publicados em 2013, dois em 2014 e um em 2015.

Os artigos selecionados na revisão de literatura apontam que existe uma combinação de recursos para a extração de relações, apresentados na seção anterior, com algoritmos de Inteligência Artificial. Constata-se ainda diferentes contextos de pesquisa entre os quais se destaca o ambiente Wikipédia (uma enciclopédia colaborativa) que envolve as próprias páginas da Wikipédia, o DBPedia (a base de dados da Wikipédia) e os *infoboxes* (as caixas de informação localizadas no lado direito das páginas da Wikipédia). Além disso, a maioria dos trabalhos encontrados aborda o problema da extração de relações no idioma inglês, mesmo naqueles em que há participação de pesquisadores brasileiros. Devido ao escopo desse artigo, apenas os trabalhos de 2014 e 2015 serão apresentados.

Paiva et al. (2014) desenvolveram um método baseado em mineração de regras de associação para o refinamento de relações de uma ontologia de domínio na área de construção civil em um processo que envolve quatro etapas, onde na primeira etapa é realizada a análise de documento; na segunda aplica-se o algoritmo *FP-Growth* para descobrir a frequência de conjuntos de termos que aparecem no vetor estatístico; na terceira etapa aplica-se regras de associação para indicar a probabilidade de um conceito ocorrer em dado outro conceito e na quarta e última etapa é realizado o mapeamento dos conjuntos de itens. Os autores utilizaram um banco de dados de regras de associação que armazena as regras definidas pelo módulo de mapeamento de conjuntos de itens frequentes. Em seguida, em uma interface são mostrados os itens, os itens candidatos e a porcentagem de frequência dos pares de itens.

Arnold e Rahm (2014) publicaram uma estratégia para criar automaticamente repositórios extraindo relações semânticas de artigos da Wikipédia em cinco etapas. A primeira é a extração de artigos da Wikipédia, com foco em cada nome e seção de resumo do artigo; a segunda etapa é o pré-processamento dos artigos extraídos; a terceira etapa identifica padrões de relações semânticas; a quarta etapa analisa os fragmentos de sentenças onde, para cada fragmento de sentença, buscam-se os conceitos relevantes ligados pelos padrões de relações semânticas. Por fim, na última etapa, há a determinação das relações semânticas. Selecionados os termos e padrões, as respectivas relações semânticas são construídas e exportadas para um repositório automaticamente.

Kliegr (2015) sugeriu a criação de um banco de dados de hiperônimos vinculados, chamado LHD (*Linked Hypernyms Dataset*), para artigos da Wikipédia nos idiomas inglês,

holandês e alemão. Para tal, quatro passos foram executados: no primeiro, um algoritmo chamado THD (*Targeted Hypernym Discovery* – Descoberta de Hiperônimos Marcados) é utilizado para análise linguística do LHD; no segundo passo, após a aplicação das características, a saída do THD é lematizada por meio do processo de retirar as flexões para determinar o seu lema; no terceiro passo determinam-se os *links* de hiperônimo realizando a desambiguação e a limpeza dos dados. Por fim, no quarto passo, realiza-se um alinhamento entre as bases DBpedia e YAGO para melhor precisão das relações definidas.

A revisão de literatura realizada foi importante para indicar as estratégias utilizadas por pesquisadores para a extração de relações semânticas. Contudo, tais estratégias não podem ser generalizadas, apesar de que elas podem ser utilizadas em parte. Desse modo, a próxima seção descreve o Semantizar, um sistema Web para desenvolvido para apoiar a extração de relações semânticas em português.

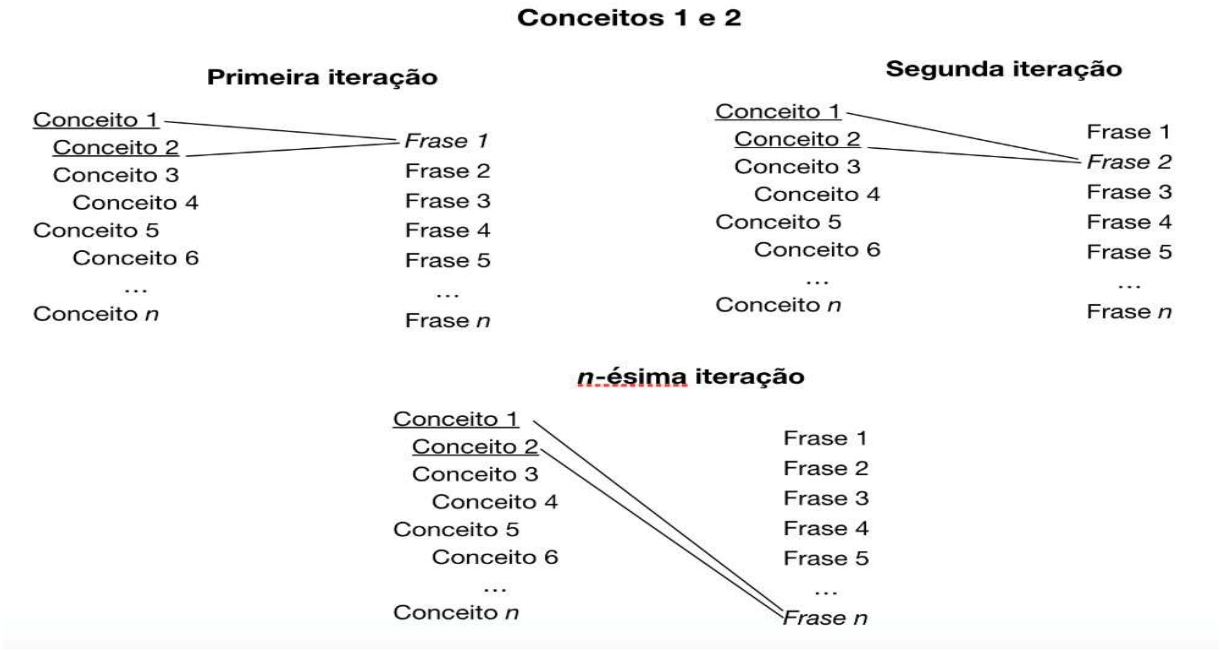
5 O SEMANTIZAR

O Semantizar é um sistema Web projetado para apoiar a extração de relações semânticas a partir de uma estrutura classificatória, originária de um documento acadêmico, e o próprio documento acadêmico. Desse modo, a estrutura classificatória contém conceitos que representam um documento acadêmico.

No Semantizar, os conceitos da estrutura classificatória são combinados uns com os outros formando pares. Para cada par de conceitos, uma busca é realizada em cada frase do documento acadêmico para verificar se eles existem na frase. A Figura 3 mostra uma representação das iterações das buscas de pares de conceitos nas frases do documento acadêmico. Como pode ser observado, no primeiro momento o par de conceitos 1 e 2 são combinados.

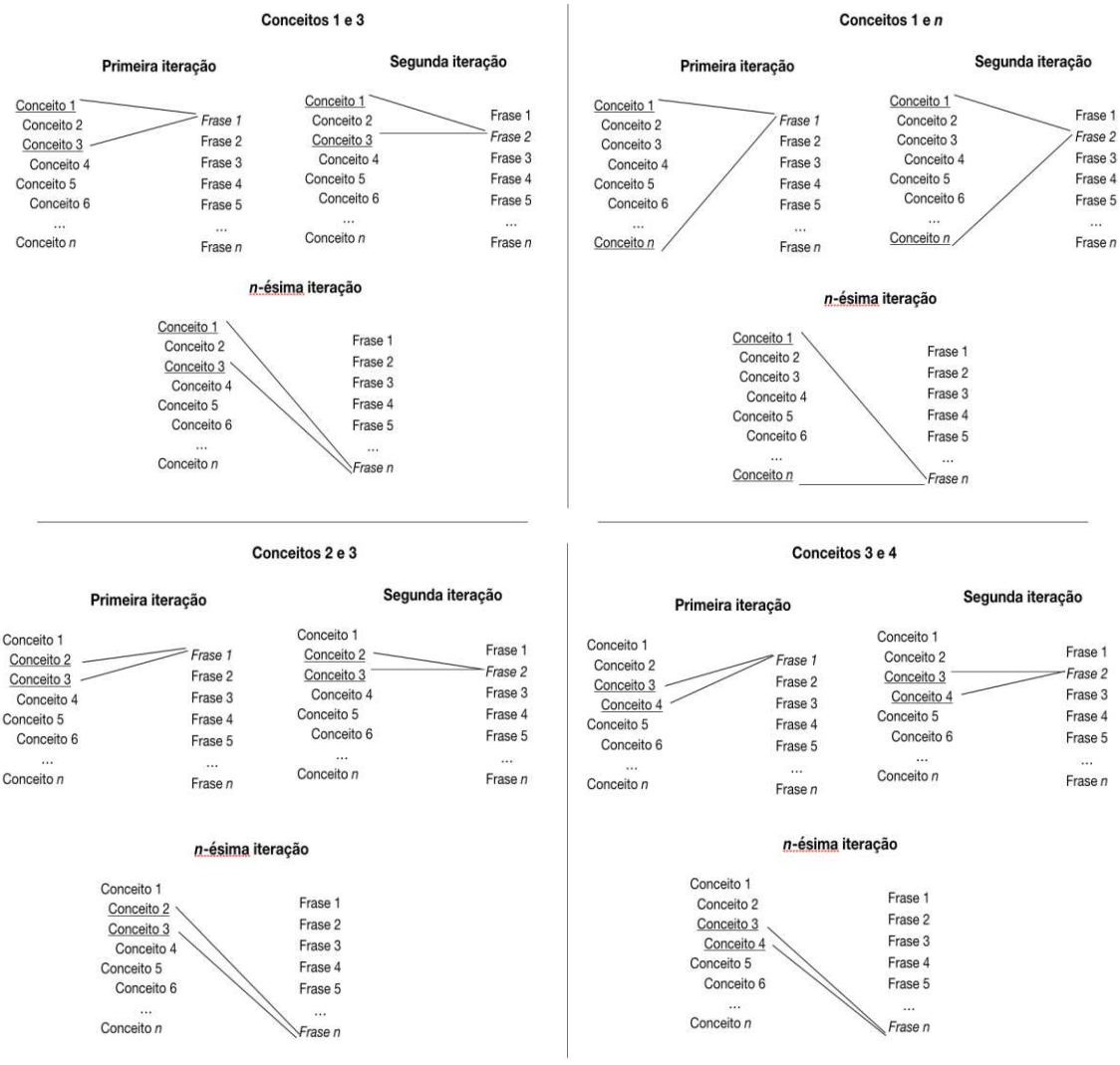
Então, examina-se a existência dessa combinação de conceitos em cada frase do documento acadêmico até a última delas. Caso o par de conceitos seja encontrado em uma frase, essa frase é destacada para que uma verificação manual confirme a existência, ou não, de uma relação semântica entre os conceitos. Se a confirmação for verdadeira, identifica-se a relação semântica encontrada entre os dois conceitos em determinada frase, indicado pela tripla: sujeito-predicado-objeto.

Figura 3: Iterações das buscas do par de conceitos 1 e 2 nas frases.



Fonte: Extraído de Maia (2018)

Figura 4: Iterações de buscas de pares de conceitos nas frases do documento acadêmico.



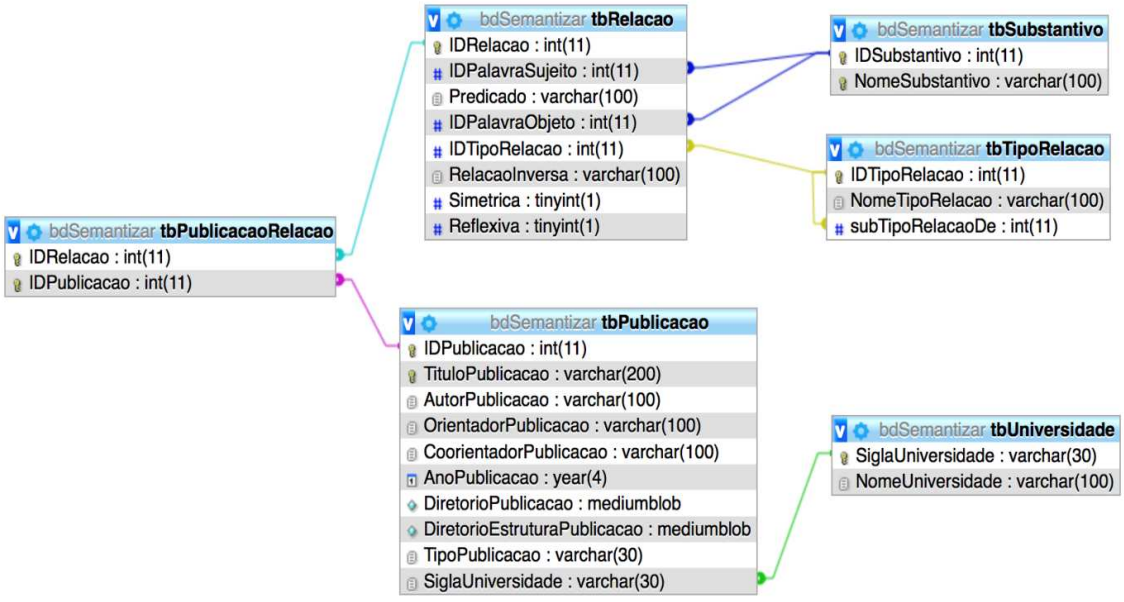
Fonte: Extraído de Maia (2018)

Em seguida, verifica-se se existe a combinação dos conceitos 1 e 3, conforme ilustrado na Figura 4. Essa verificação é realizada na primeira frase do documento acadêmico e continua até a última delas. Da mesma forma como entre os conceitos 1 e 2, explicitada anteriormente, ao se constatar os conceitos 1 e 3 em uma frase, essa frase é destacada para que seja verificada a existência de uma relação semântica entre eles. Seguidamente, o Semantizar combina o conceito 1 com todos os outros conceitos da estrutura classificatória, verificando em cada combinação se elas existem em uma frase, desde a primeira frase até a última. Ao findar as combinações com o conceito 1, o Semantizar subsequentemente faz combinações com o conceito 2 e verifica se as mesmas existem em todas as frases e assim sucessivamente até o último par de conceitos da estrutura classificatória e a última frase do documento acadêmico.

5.1 A modelagem de dados do Semantizar

O modelo de dados do Semantizar contempla seis tabelas para armazenar os dados dos substantivos, publicações, relações semânticas e relações por publicação. Além disso, foram elaboradas as tabelas para armazenar as siglas e os nomes das universidades e os tipos de relações semânticas, conforme pode ser observado no diagrama de entidade-relacionamento, na Figura 5.

Figura 5: Diagrama de Entidade Relacionamento do Semantizar.



Fonte: Extraído de Maia (2018)

Iniciando pela tabela *tbTipoRelacao,* ela armazena os tipos de relações semânticas existentes tomando como base a taxonomia proposta por Maia, Lima e Maculan (2017). Nessa taxonomia, foram levantados 63 tipos de relações semânticas classificadas como relações hierárquicas, associativas e equivalentes. Essas classes, por sua vez, têm subdivisões. Para dar suporte à taxonomia das relações semânticas, a tabela *tbTipoRelacao* contempla o autorrelacionamento por meio do campo *subTipoRelacaoDe*, que permite que um tipo de relação semântica possa ser um subtipo em outra relação. Isso pode ser observado na Figura 6, em que no tipo de relação *Hierárquica,* por exemplo, o subtipo é nulo, conforme destacado na figura. Já o tipo de relação *Objeto Estruturado* tem como subtipo a relação semântica cujo código é 19, que remete ao *Merônimo-Holônimo,* que, por sua vez, direciona para a relação *Hierárquica.*

Figura 6: Recorte da base de dados da tabela *tbTipoRelacao.*

IDTipoRelacao ▲ 1	NomeTipoRelacao	subTipoRelacaoDe
7	Hierárquica	*NULL*
8	Hipônimo-Hiperônimo	7
9	Hipônimo Simples	8
10	Instância	8
11	Inclusão de Classe	8
12	Perceptivelmente Subordinado	11
13	Funcionalmente Subordinado	11
14	Estado Subordinado	11
15	Atividade Subordinada	11
16	Geograficamente Subordinado	11
17	Taxonômica	11
18	Inclusão Espacial	8
19	Merônimo-Holônimo	7
20	Objeto Estruturado	19
21	Componente-Complexo	20
22	Unidade-Organização	20

Fonte: Extraído de Maia (2018).

Conforme apresentou-se na Figura 5, a tabela *tbsubstantivo* foi criada para armazenar os termos das estruturas classificatórias. A nomeação dessa tabela considerou que todos os termos da estrutura pertencem à classe gramatical dos substantivos; contudo, inicialmente não interessa discriminar o tipo de substantivo, a saber: comum, próprio, simples, composto, concreto, abstrato, primitivo, derivado ou coletivo. O cadastro dos registros dessa tabela é realizado automaticamente pelo sistema. Essa tabela tem a característica de ser considerada

simples por armazenar apenas o identificador do substantivo e o substantivo em si. Ressalta-se que a tabela foi planejada para restringir a duplicação de registros de substantivos.

Já a tabela *tbUniversidade* armazena o registro das principais universidades do Brasil, mas não inclui todas. Essa tabela é importante para a tabela *tbPublicacao*, que armazena, além da sigla da universidade, dados como o título da publicação (projetado de modo a não permitir duplicação); os nomes do autor, do orientador e do co-orientador (se houver); o ano da publicação; o tipo do documento, sendo permitidos um dos tipos: tese ou dissertação; e os campos que armazenam os arquivos relativos à publicação e à estrutura classificatória.

A tabela *tbrelacao* armazena a tripla sujeito-predicado-objeto, em que o sujeito e o objeto são provenientes da tabela *tbsubstantivo* e o predicado (que é a relação semântica) é definido pelo usuário. Além disso, registra-se o tipo da relação (proveniente da tabela *tbTipoRelacao*, também explicada anteriormente), a relação inversa (quando ela existir) e as propriedades relativas à simetria e reflexividade.

Por fim, a tabela *tbPublicacaoRelacao* contém campos que armazenam o código que identifica a relação semântica proveniente da tabela *tbRelacao*, explicada logo acima, e o código identificador da publicação, advindo da tabela *tbPublicacao*. Esta última é importante para associar a relação semântica ao seu contexto, que é a publicação em questão.

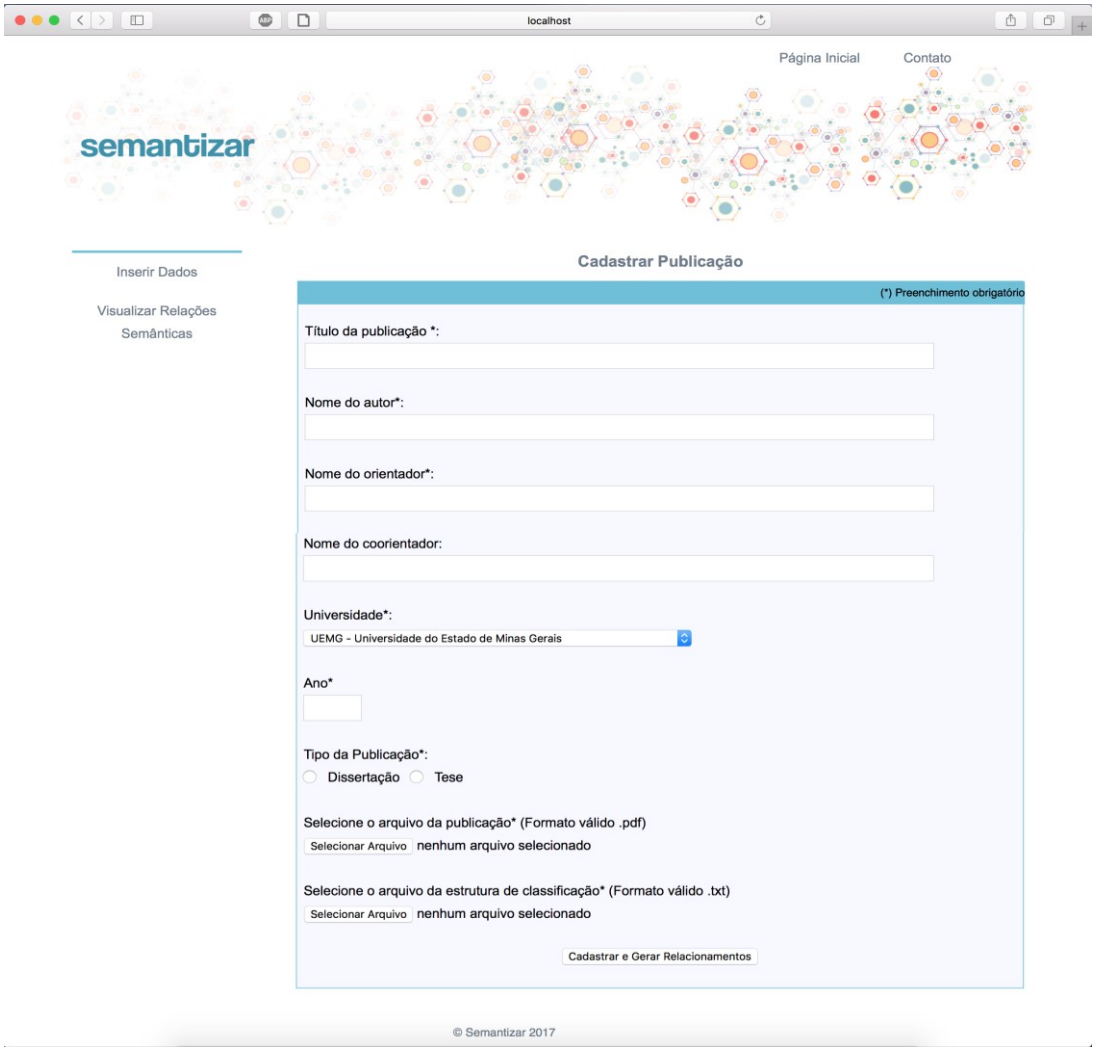
5.2 A implementação do Semantizar

A implementação do Semantizar foi dividida em quatro atividades: (1) entrada de dados, (2) leitura e preparação, (3) extração de relações semânticas e (4) representação do conhecimento.

A primeira atividade, a de entrada de dados, é responsável por receber os metadados do documento acadêmico do qual se pretende extrair relações semânticas. Os metadados incluem o título da publicação, o nome do(a) autor(a), o nome do(a) orientador(a) e, se houver, o nome do(a) co-orientador(a), o local de publicação, o ano em que o documento foi defendido e, por fim, o tipo do documento acadêmico, se tese ou dissertação. Além disso, na entrada de dados, envia-se os arquivos referentes à publicação e à estrutura classificatória, sendo que a publicação é um arquivo do tipo .pdf (*Portable Document Format* – Formato Portátil de Documento) e a estrutura classificatória é um arquivo de texto simples de extensão .txt. A Figura 7 mostra a interface de cadastro implementada para esta etapa do Semantizar.

A atividade subsequente, de leitura e preparação, verifica se cada termo da estrutura classificatória existe na base de dados da seguinte forma: considerando que a linguagem de programação PHP (*Hypertext Preprocessor*), escolhida para a implementação do Semantizar, permite que arquivos de texto sejam convertidos em vetores, o arquivo referente à estrutura classificatória é convertido automaticamente em um vetor de termos. Cada linha da estrutura (que se refere a um termo) é transformada em uma posição do vetor. Logo, o algoritmo percorre cada posição do vetor verificando se o conteúdo da posição, que é o termo da estrutura classificatória, existe na tabela tbSubstantivo no banco de dados; caso não exista, o termo é automaticamente cadastrado pelo sistema. A outra tarefa da atividade de leitura e preparação é a preparação do arquivo referente à publicação para a manipulação que ocorrerá na próxima etapa do Semantizar que consiste em converter o documento acadêmico em formato .pdf para um arquivo temporário no formato .txt.

Figura 7: Interface da atividade de entrada de dados do Semantizar.



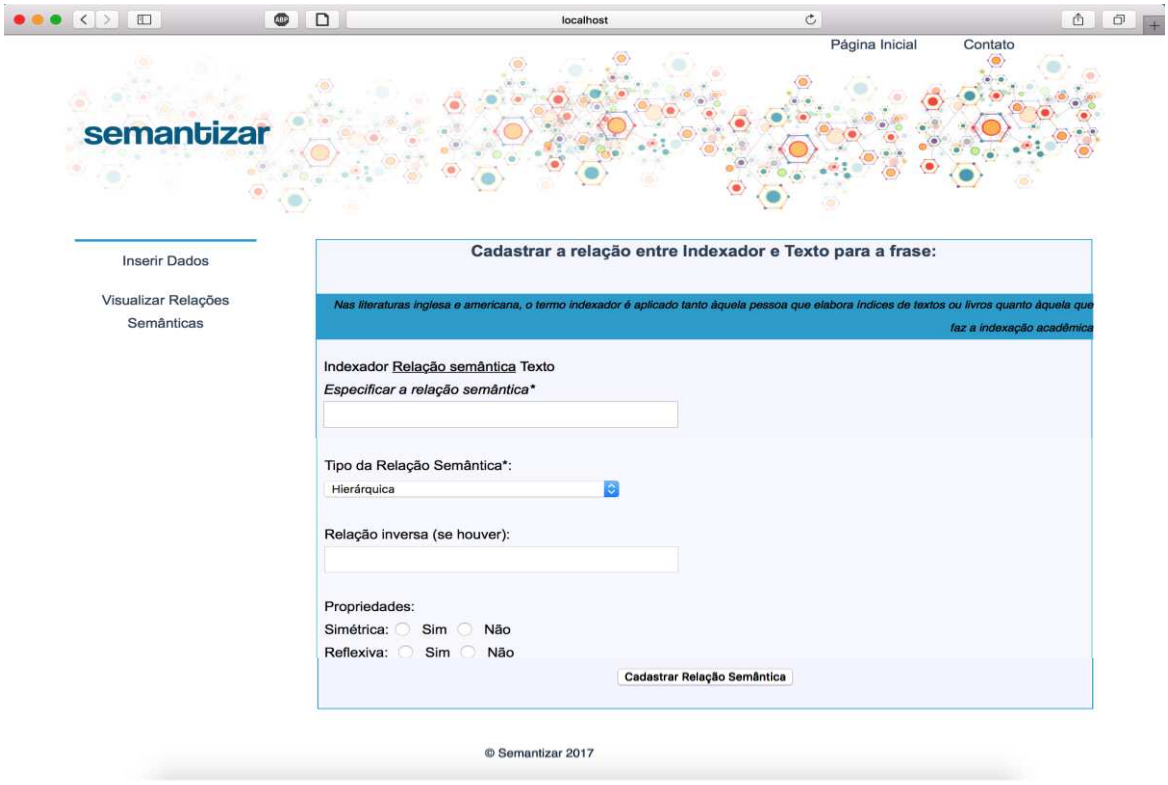
Fonte: Extraído de Maia (2018).

A atividade de extração de relações semânticas é considerada a mais importante e o cerne do Semantizar. Ela é composta de duas tarefas, sendo que na primeira realiza-se uma busca por pares de termos da estrutura classificatória em frases da publicação à qual a estrutura se refere. O arquivo temporário criado na etapa anterior é transformado em uma *string* (variável que armazena caracteres alfanuméricos). Posteriormente, essa *string* é decomposta em uma *string* menor (que denota uma frase) toda vez que o sinal de pontuação ponto-final (.) é encontrado no arquivo, criando um vetor de frases. Em seguida, o algoritmo percorre o vetor de frases e verifica se existe em cada posição desse vetor o termo do vetor de termos. Caso exista, ele percorre as demais posições desse mesmo vetor de termos para verificar se existe outro termo da estrutura na mesma frase. Se existir, um formulário é criado para o usuário validar a relação semântica. A validação da relação semântica é a segunda tarefa da atividade de extração de relações. A Figura 8 mostra a interface implementada para essa fase. Caso o usuário concorde que existe uma relação semântica entre os dois conceitos da estrutura classificatória encontrados em uma frase da publicação, ele é submetido a outro formulário, apresentado na Figura 9, em que a relação semântica é cadastrada.

Figura 8: Interface de validação das relações semânticas.

Fonte: Extraído de Maia (2018).

Figura 9: Interface de cadastro da relação semântica.



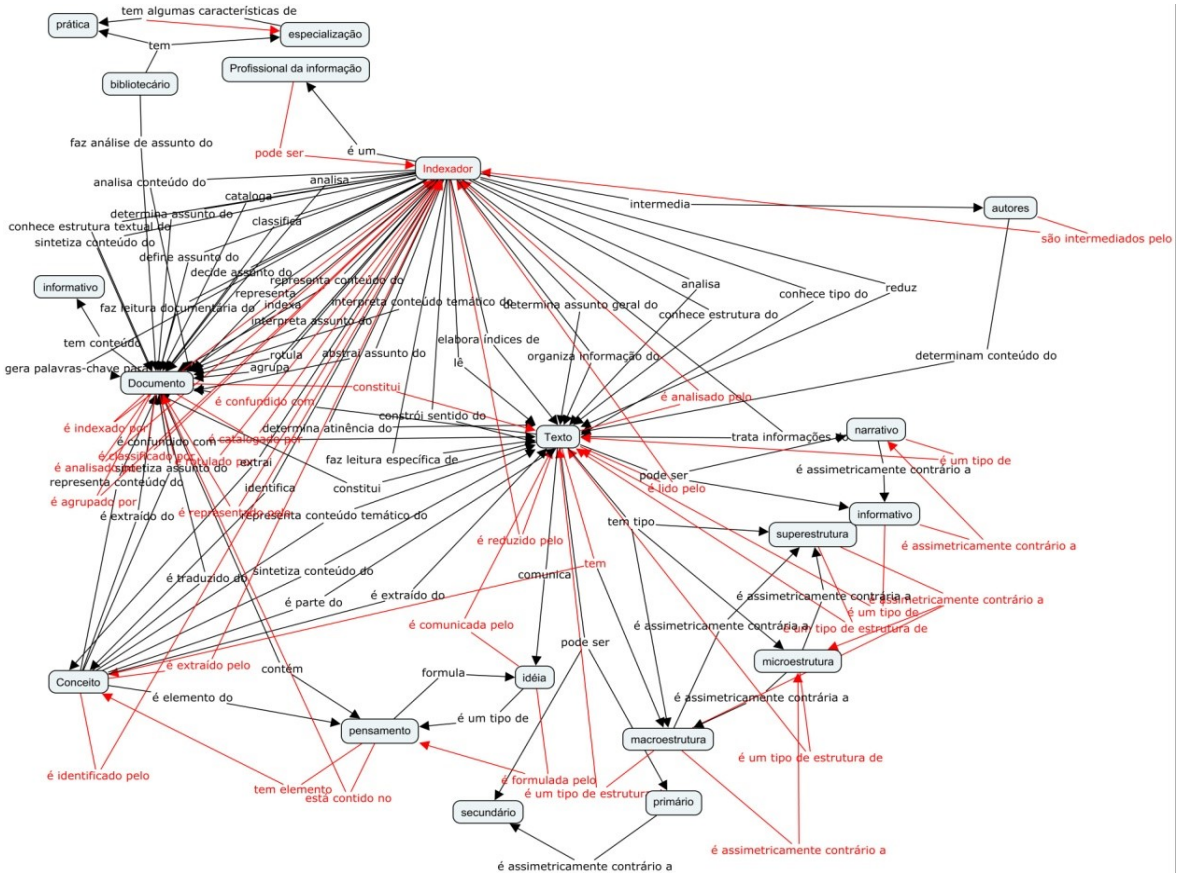
Fonte: Extraído de Maia (2018).

A interface de cadastro apresentada na Figura 9 é criada para cada indício de existência de relação semântica, considerando que, quando existem dois conceitos da estrutura classificatória na mesma frase, pode haver a ocorrência de uma relação entre esses conceitos. O usuário pode especificar a relação semântica conforme seu julgamento ao analisar a frase (por exemplo, *elabora índices de,* conforme a análise do trecho da publicação apresentado na Figura 9). Em seguida, o usuário especifica qual o tipo da relação semântica. Conforme mencionado na seção 5.1, os tipos de relações semânticas consideraram a taxonomia de relações semânticas elaborada por Maia, Lima e Maculan (2017). Posteriormente, o usuário informa a relação inversa, se houver. No caso do exemplo apresentado, a seguinte relação inversa só seria possível se o termo da estrutura classificatória fosse *índices de texto* e não *texto***.** Nesse caso, a tripla de relação semântica inversa seria: Índices de texto *são elaborados pelo* indexador. Logo, no campo relação inversa, o usuário informaria: "são elaborados pelo". Finalizando o cadastro das relações semânticas, o usuário assinala as propriedades: simetria e reflexividade.

6 AVALIAÇÃO DO SEMANTIZAR

Esta seção apresenta os resultados de um estudo de caso realizado para examinar as contribuições do Semantizar como facilitador da explicitação de relações semânticas. Após a execução do recorte da amostra no Semantizar, a validação, análise e interpretação dos dados, resultados importantes foram levantados. O primeiro deles é o mapa conceitual com a compilação das relações semânticas explicitadas (Figura 10). Nesse mapa, as relações semânticas destacadas na cor vermelha são relações inversas.

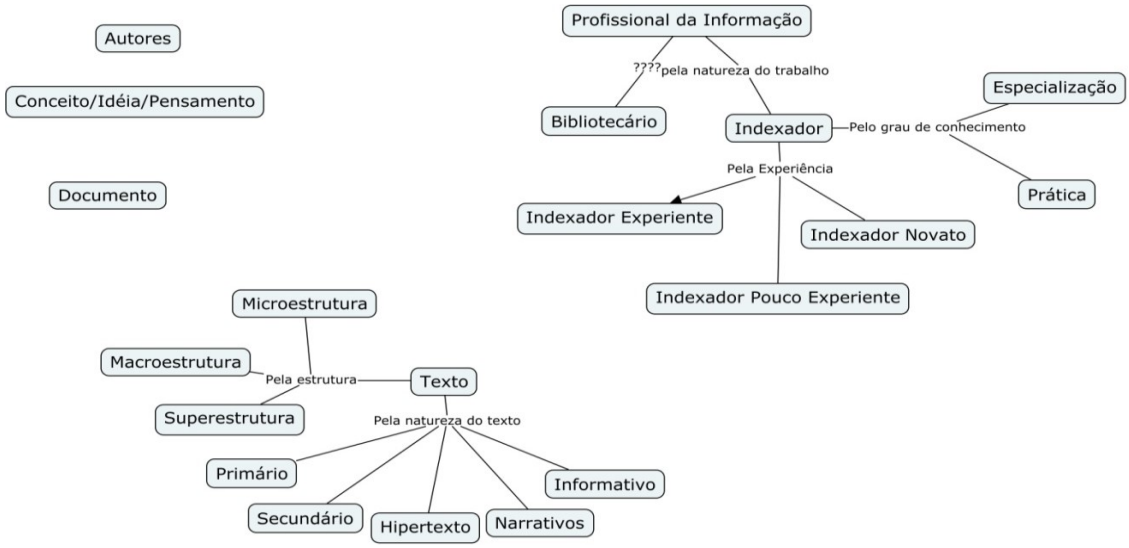
Figura 10: Mapa conceitual das relações semânticas explicitadas no estudo de caso.



Fonte: Extraído de Maia (2018)

Já a Figura 11 é um mapa conceitual gerado a partir da estrutura classificatória sem o intermédio do Semantizar. Ao observar essa Figura constata-se que as relações semânticas explicitadas são decorrentes da nomeação das subfacetas utilizadas por Lima (2004) (Veja as subfacetas em destaque na Figura 12). Observa-se também que existe uma relação semântica entre *profissional da informação* e *bibliotecário*; inerente da indentação que denota um tipo de hierarquia. Contudo pela estrutura facetada não foi possível especificar qual é de fato a relação semântica.

Figura 11: Mapa conceitual da estrutura classificatória.



Fonte: Extraído de Maia (2018).

Figura 12: Estrutura facetada utilizada na amostra.

Personalidade [Entities]

- Autores
- Profissional da informação
  - Bibliotecário

(Pela natureza do seu trabalho)

  - Indexador

(Pela experiência)

    - Indexador experiente
    - Indexador pouco experiente
    - Indexador novato

(Pelo grau de conhecimento)

    - Especialização
    - Prática
- Conceito/Idéia/Pensamento
- Documento
- Texto

(Pela natureza do texto)

  - Narrativos
  - Informativo
  - Primário
  - Secundário
  - Hipertexto

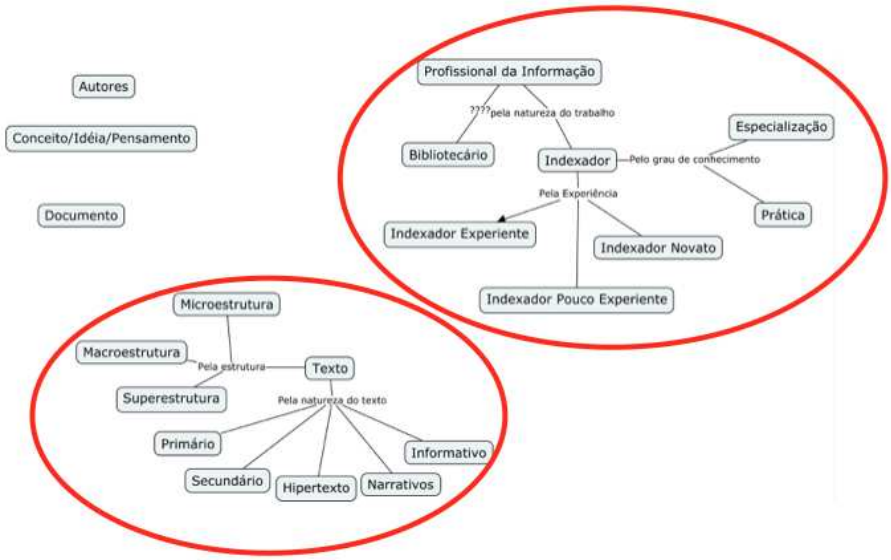
(Pela estrutura)

  - Microestrutura
  - Macroestrutura
  - Superestrutura

Fonte: Gercina Lima (2004)

Na organização da estrutura facetada no mapa conceitual, verifica-se a presença de dois *clusters*, conforme destacado na Figura 13. O primeiro grupo é composto por conceitos relacionados a *texto* e o outro a *indexador*.

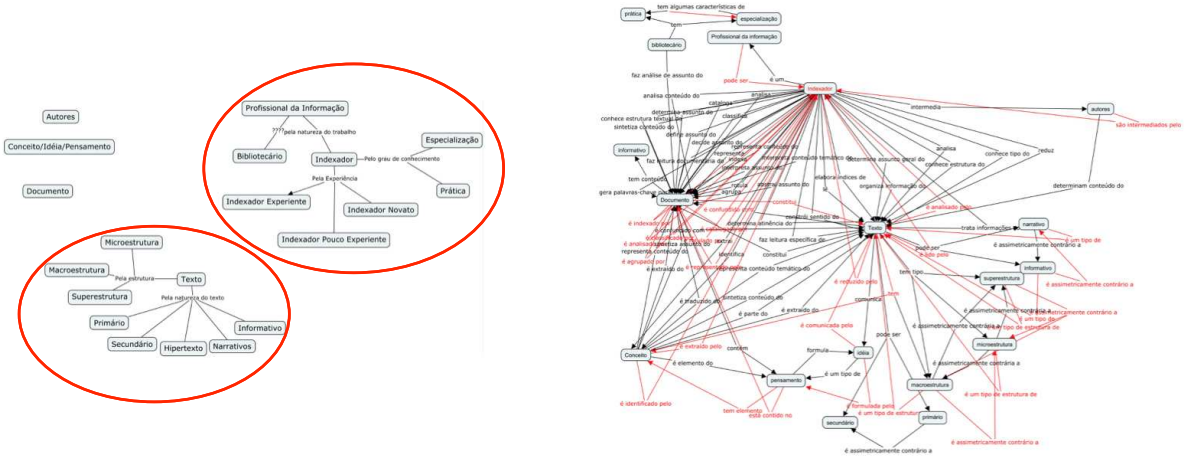
Figura 13: *Clusters* resultantes do mapa conceitual da estrutura classificatória.



Fonte: Extraído de Maia (2018)

Nos *clusters* os objetos pertencentes a um grupo são relacionados entre si, porém, eles não se relacionam com os conceitos que estão fora de seu grupo. Portanto, como pode ser visto na Figura 14, no mapa conceitual gerado a partir das relações semânticas encontradas no Semantizar, há uma coesão entre todos os conceitos de tal maneira que esses *clusters* não são possíveis de serem obtidos, ou seja, os conceitos relacionam-se todos entre si.

Figura 14: Mapa conceitual gerado sem/com as relações semânticas extraídas pelo Semantizar.



Mapa conceitual criado **sem** as relações semânticas extraídas pelo Semantizar

Mapa conceitual criado **com** as relações semânticas extraídas pelo Semantizar

Fonte: Extraído de Maia (2018).

Os relacionamentos entre todos os conceitos só foram possíveis com o apoio do Semantizar, que permitiu a criação de uma representação que cobre todos os conceitos de uma

estrutura estrutura classificatória semanticamente relacionados de tal forma que o usuário consiga vislumbrar todas as relações possíveis entre os conceitos. Nesse sentido, há um ganho de informação para o usuário. Desse modo, pode-se afirmar que o Semantizar enriqueceu semanticamente uma estrutura classificatória.

6.1 Contribuições do Semantizar

O enriquecimento semântico de estruturas classificatórias para a representação do conhecimento realizado manualmente pode demandar tempo e esforço por parte do profissional que o realiza. Portanto, o Semantizar facilitou a extração e a explicitação de relações semânticas essenciais para a representação do conhecimento, semiautomatizando essa tarefa. Desse modo, o Semantizar contribuiu para a representação do conhecimento a partir de uma estrutura classificatória, mostrando-se objetivo ao detectar dois conceitos em uma frase de um texto extenso. Entende-se que essa identificação dos pares de conceitos em cada frase é uma das etapas mais laboriosa no contexto que o Semantizar foi criado para atuar.

Por meio do Semantizar, foi possível explicitar 101 relações semânticas, incluindo as inversas, em 53 diferentes pares de conceitos, também considerando as relações inversas, a partir de uma estrutura classificatória contendo 22 conceitos. Com isso, foi possível melhorar a semântica da amostra.

Outra contribuição importante do Semantizar foi que, de certa forma, o Semantizar atuou como um agente de validação da estrutura classificatória, pois ele indicou 199 indícios de relações semânticas em 131 frases do recorte do documento acadêmico ao qual a estrutura classificatória representou. Acredita-se que essa quantidade de indícios é um indicador importante para sugerir que a estrutura relevantemente representa o documento acadêmico.

Além disso, foi possível, por meio do Semantizar, indicar os pares de conceitos mais importantes para o documento acadêmico. Verificou-se durante a análise dos dados que os pares de conceitos que mais ocorreram denotaram ser os mais relevantes no domínio representado. No caso da amostra, destacou-se os pares: *indexador* e *documento* e *indexador* e *texto*. Tomando como referência que esses pares foram extraídos da tese intitulada *Fatores interferentes no processo de análise de assunto: estudo de caso de indexadores*, de Naves (2000), naturalmente pode-se dizer, mesmo sem ler todo o conteúdo do documento acadêmico, que tais pares são os mais importantes do domínio. Consequentemente, o Semantizar também

indicou os conceitos mais importantes para o documento acadêmico. Do mesmo modo como para os pares de conceitos, tomando como referência os conceitos que mais ocorreram nas relações semânticas, pode-se afirmar que os conceitos mais importantes da tese de Naves (2000) foram *indexador*, *texto* e *documento*, sendo que *texto* e *documento* foram classificados como quase sinônimos, o que também indica coerência na determinação dessa relação semântica.

Por fim, notou-se, que a extração dos conceitos pode ser realizada a partir de listas de termos, pois a hierarquia inerente da estrutura classificatória não foi determinante no Semantizar para a indicação da existência de uma relação semântica.

7 CONSIDERAÇÕES FINAIS

Os documentos acadêmicos, como as dissertações e teses, são fontes de conhecimento explícito. Contudo, algumas vezes, tal conhecimento pode ser de difícil compreensão devido à complexidade que envolve algumas pesquisas. Desse modo, a representação do conhecimento pode facilitar o entendimento desses documentos. Na representação do conhecimento, as relações semânticas entre os conceitos contribuem com o enriquecimento semântico do domínio a partir de (ou compondo) bases conceituais, como as estruturas classificatórias, uma vez que as relações semânticas permitem a compreensão da natureza que envolve a ligação entre os conceitos em determinado contexto.

Nesse sentido, o Semantizar mostrou que é possível apoiar a representação do conhecimento de textos em linguagem natural sugerindo relações semânticas entre os conceitos em seus contextos. Um mapa conceitual com todas as relações entre os conceitos revelou relações semânticas que *a priori* a estrutura classificatória não sustentava.

Ainda, constatou-se que o Semantizar contribuiu com estudos sobre a extração de relações semânticas em português brasileiro, pois, com base na literatura pesquisada, este trabalho é pioneiro na extração de relações a partir de uma estrutura classificatória, provendo um objeto de estudos futuros para a extração automática de relações semânticas em português brasileiro.

REFERÊNCIAS

ARNOLD, P.; RAHM, E. Extracting Semantic Concept Relations from Wikipedia. **4th International Conference on Web Intelligence**, Mining and Semantics, 2014.

GRISHMAN, R.; SUNDHEIM, B. **Message Understanding Conference-6:** A Brief History. [S.l: s.n.], 1996, p. 466-471.

KLIEGR, T. Linked hypernyms: Enriching DBpedia with Targeted Hypernym Discovery. **Journal of Web Semantics**, [S.l.], v. 31, p. 59-69, mar. 2015.

KUMAR, E. **Natural Language Processing**. [S.l.]: I. K. International Pvt Ltd, 2011.

LIMA, G. Â. B. DE O. **MAPA HIPERTEXTUAL (MHTX):** Um modelo para organização hipertextual de documento. 2004. 207 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2004.

MAIA, L. S. **Extração e explicitação de relações semânticas para a representação do conhecimento de documentos acadêmicos:** um estudo de caso a partir de uma estrutura classificatória. 2018. 248f. Tese (Doutorado) - Universidade Federal de Minas Gerais, Belo Horizonte, 2018.

MAIA, L. S.; LIMA, G. A. ; MACULAN, B. C. M. S. Taxonomia dos tipos de relações semânticas para a Organização e Representação do Conhecimento: Uma proposta a partir da literatura. ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO. 18., 2017. **Anais do ENANCIB**, Marília, SP, 2017,

NAVES, M. M. L. **Fatores interferentes no processo de análise de assunto**: estudo de caso de indexadores. 2000. 283 f. Tese (Doutorado) – Universidade Federal de Minas Gerais, Belo Horizonte, 2000.

PAIVA, L. *et al.* Discovering Semantic Relations from Unstructured Data for Ontology Enrichment Asssociation rules based approach*. In:* ROCHA, A. *et al.* (Org.). **Proceedings of the 2014 9th Iberian Conference on Information Systems and Technologies** *(cisti 2014)*. New York: Ieee, 2014.

PRADONOV, C. C.; FREITAS, E. C. DE. *Metodologia do Trabalho Científico:* Métodos e Técnicas da Pesquisa e do Trabalho Acadêmico. 2, ed. Novo Hamburgo, Rio Grande do Sul: Editora Feevale, 2013.

SARAWAGI, S. Information extraction***. Foundations and Trends® in Databases***, [S.l.], v. 1, n. 3, p. 261-377, 2008.

SILVEIRA, D. T.; CÓRDOVA, F. P. **A pesquisa científica**. Métodos de pesquisa. Porto Alegre: Editora da UFRGS, 2009, p. 31-42.

WOHLIN, C. *et al.* **Experimentation in Software Engineering**. [S.l.]: Kluwer Academic Publishers Boston/Dordrecht/London, 2000.