Pontifícia Universidade Católica do Rio Grande do Sul
Faculdade de Informática
Pós-Graduação em Ciência da Computação

SiSe: Medida de similaridade semântica entre ontologias em português

Juliano Baldez de Freitas

Dissertação apresentada como requisito parcial à obtenção do grau de mestre em Ciência da Computação

Orientadora: Profa. Dra. Vera Lúcia Strube de Lima

Porto Alegre, agosto de 2007

PUCRS
Campus Central
Av. Ipiranga, 6681 - prédio 16 - CEP 90619-900
Porto Alegre - RS - Brasil
Fone: +55 (51) 3320-3544 - Fax: +55 (51) 3320-3548
Email: bceadm@pucrs.br
www.pucrs.br/biblioteca

Pontifícia Universidade Católica do Rio Grande do Sul
FACULDADE DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

TERMO DE APRESENTAÇÃO DE DISSERTAÇÃO DE MESTRADO

Dissertação intitulada ***"SiSe: Medida de Similaridade Semântica entre Ontologias em Português"***, apresentada por Juliano Baldez de Freitas, como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 26/01/2007 pela Comissão Examinadora:

Profa. Dra. Vera Lúcia Strube de Lima – Orientadora — PPGCC/PUCRS

Prof. Dr. Marcelo Blois Ribeiro – — PPGCC/PUCRS

Prof. Dr. Marco Antonio Insaurriaga Gonzalez – — FACIN/PUCRS

p/ Profa. Dra. Aline Villavicencio – — UFRGS

Homologada em 09/08/07, conforme Ata No. 18 pela Comissão Coordenadora.

Prof. Dr. Fernando Luís Dotti
Coordenador.

PUCRS
Campus Central
Av. Ipiranga, 6681 – P. 16 – sala 106 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320-3621
E-mail: ppgcc@inf.pucrs.br
www.pucrs.br/facin/pos

Dedico esta dissertação a minha família, meus pais Alceu e Denisia e meu irmão Alceu Júnior.

Agradecimentos

Primeiramente agradeço a meus pais, Alceu e Denisia, por me darem todo o apoio necessário para que eu pudesse aprofundar meus estudos em uma pós-graduação em nível de mestrado, bem como o incentivo dado durante toda a vida acadêmica e pessoal. Também, a meu irmão Alceu Júnior pelo apoio.

À minha namorada Vanessa, que me apoiou e entendeu os momentos de ausência.

À minha orientadora, Profa. Vera que foi muito compreensiva e muito presente durante estes dois anos do mestrado. Obrigado pelos ensinamentos, dicas e incentivo dados. Também, por proporcionar o mestrado-sanduíche para USP de São Carlos e viagens para congressos, as quais foram muito enriquecedoras profissionalmente.

Agradeço, também aos professores Marco Ganzalez e Marcelo Blois pelas dicas e apoio durante os trabalhos individuais, seminário de andamento e PEP. Em especial ao Prof. Marcelo Blois como orientador do estágio de docência. Também, a todo corpo docente do PPGCC que direta ou indiretamente contribuíram para minha formação neste mestrado.

Ao grupo de pesquisa do NILC que me recebeu de forma extraordinária na USP de São Carlos, durante os dois meses de Mestrado-Sanduíche. Especialmente a Profa. Maria das Graças Volpe Nunes encarregada de me orientar neste período.

Ao pessoal da república na qual me hospedei durante os dois meses em São Carlos: Ricardo Hasegawa, Márcio, Reginaldo Ré. Obrigado pelas dicas de São Carlos, pelos churrascos à moda paulista e pela amizade.

Agradeço também a Marcirio Chaves que apesar de estar em Portugal em seu doutorado contribuiu diretamente na minha dissertação com dicas e sugestões sempre atualizadas de referências bibliográficas, respondendo os e-mails com muita rapidez e boa vontade :-).

Ao CDPe - Centro de Desenvolvimento e Pesquisa Dell-PUCRS pelo financiamento da bolsa de mestrado.

A todos os colegas do CDPe com os quais convivi neste período do mestrado.

Aos amigos Marcos Cardoso e Maria Isabel, pelas manhãs, tardes e noites de estudo que fizemos juntos neste mestrado, e pela amizade desde os tempos da graduação na UCPel.

Aos meus grandes amigos e colegas de apartamento que passaram pelo 338 nestes dois anos: Pablo Grigoletti, Eduardo Bastos e Cristovão Pereira. Obrigado pela amizade, companheirismo e pelas horas de descontração proporcionadas.

Aos amigos, Eduardo Menna e Diego Menna pela amizade de vários anos, me apoiando e incentivando durante o mestrado.

Resumo

Este trabalho consiste na adaptação de uma medida de similaridade semântica para o mapeamento entre ontologias em português. A medida SiSe (Similaridade Semântica) apresentada neste trabalho adapta a proposta Mapeamento Taxonômico, de Maedche e Staab [Maedche e Staab 2002].

A medida SiSe faz uma comparação da similaridade entre termos de ontologias distintas através da análise da hierarquia dos mesmos. Utilizamos o conceito de "*Semantic Cotopy*" e "*Common Semantic Cotopy*", os quais formam um conjunto para cada um dos termos comparados. Cada conjunto é composto pelo termo, pelos subconceitos e superconceitos deste termo, todos representados por seus *stems*, através de um recurso de Processamento da Linguagem Natural, o *stemmer* PortugueseStemmer desenvolvido por Orengo e Huyck [Orengo e Huyck 2001].

Nossa medida adota uma estratégia para o mapeamento entre ontologias que envolve a análise das linguagens utilizadas na descrição das ontologias (OWL, RFDS, etc), abstraindo as sintaxes e normalizando em uma linguagem XML com as relações hierárquicas de hiponímia e hiperonímia das ontologias. A medida de similaridade SiSe compara as ontologias através das relações hierárquicas que as mesmas possuem, desta forma o coeficiente resultante é a similaridade semântico-estrutural entre os termos das ontologias.

A avaliação da medida SiSe é realizada através de um "*Golden mapping*", ou mapeamento dourado, que consiste na avaliação da similaridade de algumas ontologias por humanos confrontando com os resultados da medida SiSe.

Esta medida é utilizada para auxiliar no mapeamento entre ontologias visando o reuso e a integração de informação.

Palavras-chave: medida de similaridade semântica, mapeamento entre ontologias, ontologias, Processamento da Linguagem Natural.

Abstract

This work concerns the development of a semantic similarity measure for mapping between Portuguese ontologies. The SiSe (*Similaridade Semântica*) measure presented in this work is an extension of the proposal known as Taxonomic Overlap proposed by Maedche and Staab [Maedche e Staab 2002].

SiSe makes a comparison on the similarity between terms of distinct ontologies through the analysis of their hierarchies. We use the concepts of "Semantic Cotopy" and "Common Semantic Cotopy", which build a set for each term in question. This set is composed by the term and the subconcepts and superconcepts of this term, all represented by their stems, through the stemmer PortugueseStemmer by Orengo and Huyck [Orengo e Huyck 2001].

Our measure adopts a mapping that considers the languages used in the description of the ontologies (for example, OWL, RFDS, etc), and normalizes them in XML keeping the hierarchic relations of hyponym and hypernym in the ontologies. The SiSe measure compares the similarity between the ontologies through the hierarchic relations that are common among them, and the result is a semantic-structural similarity value.

The evaluation of the SiSe measure is carried out through a "Golden mapping" that consists of the similarity between two ontologies according to human analysis. The SiSe results are compared to this Golden Mapping.

This measure helps the mapping between ontologies aiming at the reuse and the information integration.

Keywords: semantic similarity measure, mapping between ontologies, ontologies, Natural Language Processing.

Lista de Tabelas

Lista de Figuras

Lista de Siglas

IA	Inteligência Artificial	18
SE	Sistemas Especialistas	18
W3C	*World Wide Web Consortium*	27
RDF	*Resource Description Framework*	27
RDFS	RDF *Schema*	27
XML	*eXtensible Markup Language*	27
URI	*Uniform Resource Identifier*	27
OWL	*Web Ontology Language*	28
DE	Distância de Edição	32
CC	Combinação de Caracteres	32
SL	Similaridade Lexical	33
FCA-Merge	*Formal Concept Analysis - Merge*	35
PLN	Processamento da Linguagem Natural	35
MT	Mapeamento Taxonômico	37
SC	*Semantic Cotopy*	37
CSC	*Common Semantic Cotopy*	38
CS	Combinação Semântica	38
SAT	*Boolean Satisfiability*	39
MAFRA	*Mapping Framework for Distributed Ontologies*	42
SiSe	Similaridade Semântica	55
TML	*Thesaurus Markup Language*	62
GM	*Golden Mapping*	67

Sumário

Capítulo 1

Introdução

1.1 Motivação e contexto do trabalho

Atualmente o uso e reuso do conhecimento é fundamental, devido à crescente quantidade de informação que está sendo gerada. O mesmo se reflete na quantidade e velocidade de geração de conhecimento. Este fato faz com que pessoas e organizações tenham que gerenciar seu conhecimento de modo mais eficaz. Combinar conhecimentos de domínios distintos pode acarretar problemas como, por exemplo, formatos de representação do conhecimento distintos e inconsistências semânticas, entre outros [Ding e Foo 2002]. Segundo Chaves em [Chaves 2003], as inconsistências semânticas são geradas quando dois sistemas projetados independentemente são integrados, ocorrendo alguns "mal entendidos" entre os significados e as interpretações para o mesmo dado como, por exemplo, nomes, estruturas, esquemas ou atributos, entre outros. Estas inconsistências semânticas também se aplicam à área de Engenharia Ontológica.

As ontologias podem ser construídas a partir de outras já existentes. Por exemplo, o autor Euzenat em [Euzenat *et al.* 2004] relata que ontologias interdisciplinares podem ser criadas a partir de ontologias de domínios específicos. Para Ding e Foo [Ding e Foo 2002], as ontologias são mecanismos para a comunicação entre agentes de software. Se estas ontologias forem incompatíveis do ponto de vista da comunicação, a troca de informações entre os agentes não é possível.

A integração ou determinação de equivalência das informações entre duas estruturas ontológicas depende do mapeamento feito entre os termos destas duas ontologias. Para Maedche e Staab em [Maedche e Staab 2002], mapear termos de duas ou mais ontologias é associar conceitos equivalentes entre elas, de acordo com uma relação de similaridade. Vários trabalhos encontrados na literatura como, por exemplo, [Euzenat *et al.* 2004], [Pinto, Gómez-Pérez e Martins 1999], [Stumme *et al.* 2000], classificam as abordagens para mapear ontologias em união e alinhamento. Ambas as abordagens utilizam uma medida de similaridade para mapear os elementos entre estruturas ontológicas.

De acordo com a literatura estudada, podemos classificar em dois tipos as medidas de similaridade usadas para mapear elementos entre ontologias: similaridade lexical e similaridade semântica. As medidas de similaridade lexical estabelecem, em geral, uma relação

entre os termos, comparando as cadeias de caracteres que constituem os termos em questão. Já as medidas de similaridade semântica se concentram no significado dos termos usados nas ontologias, buscando correspondências através do sentido que estes termos expressam. A similaridade semântica leva em consideração o significado e as relações semânticas existentes entre os elementos de duas estruturas ontológicas distintas envolvidos no processo de medição da similaridade, bem como as relações semânticas que cada elemento possui na sua estrutura ontológica, [Euzenat *et al.* 2004], [Maedche e Staab 2002] e [Giunchiglia, Yatskevich e Giunchiglia 2005].

O presente trabalho tem como objetivo o estudo de abordagens de identificação (ou cálculo) de similaridade lexical e semântica existentes na literatura, para propôr uma medida de similaridade semântica entre ontologias em português.

1.2 Organização do texto da dissertação

Para facilitar o entendimento e a contextualização dos conceitos envolvidos no presente trabalho, foi adotada a seguinte estruturação do texto:

- **Capítulo 2**: aborda os conceitos de ontologias, bem como o mapeamento entre ontologias e as abordagens para realizar este mapeamento (união e alinhamento). São apresentados, brevemente, os tipos de relações semânticas existentes entre palavras, bem como algumas linguagens para construção de ontologias (RDF/RDFS e OWL).

- **Capítulo 3**: apresenta algumas das abordagens encontradas na literatura para medir a similaridade lexical e semântica entre elementos de ontologias distintas.

- **Capítulo 4**: este capítulo apresenta os conceitos da medida Mapeamento Taxonômico de Maedche e Staab, a qual é adaptada para o português através da medida SiSe. São apresentados exemplos de cálculo da similaridade, adaptações e estratégias para o mapeamento entre ontologias usando a medida SiSe. Também são apresentadas as características de um protótipo desenvolvido para facilitar o mapeamento e análise dos resultados dos experimentos.

- **Capítulo 5**: neste capítulo é descrito o método usado para avaliação dos resultados da medida SiSe. São descritos os passos executados para a avaliação, as ontologias utilizadas e a análise dos resultados.

- **Capítulo 6**: traz as considerações finais quanto ao trabalho realizado. Traz, também, possíveis temas de trabalhos futuros a partir deste estudo.

Capítulo 2

Fundamentação teórica

2.1 Ontologias na Ciência da Computação

A palavra "ontologia" tem origem grega: foi introduzida primeiramente na Filosofia, por Aristóteles. A abordagem filosófica sobre ontologias especifica o que existe no mundo e o que podemos dizer sobre o mundo, ou seja, demonstra uma teoria sobre a natureza e a existência do ser [Gruninger e Lee 2002].

Aristóteles observou que o conhecimento existente no mundo poderia ser representado através de objetos, e que os mesmos poderiam ser classificados. Os objetos classificados herdariam as características da classe a que fossem inseridos. Estas classes deveriam estar em uma certa ordem, de acordo com uma estrutura para construção de hierarquias de classes baseadas em conceitos de alto nível, criadas por Aristóteles, conhecidas como *categorias* - por exemplo: *substância, quantidade, relação, lugar, tempo*, etc. Estas abstrações utilizadas para definir conhecimento de domínio através de classes, criadas por Aristóteles há séculos atrás, formaram a base de muitos conceitos atualmente utilizados na Ciência da Computação como, por exemplo, orientação a objetos, raciocínio baseado no senso comum e ontologias [Freitas, Stuckenschmidt e Noy 2005].

Na Computação, o uso do termo ontologia teve origem na comunidade da Inteligência Artificial (IA) [Maedche 2002]. As ontologias foram muito utilizadas em sistemas baseados em conhecimento como, por exemplo, em Sistemas Especialistas (SE) na década de oitenta, pois fornecem uma estrutura declarativa a qual é utilizada na tentativa de inferência automática. Desde então, o uso e construção de ontologias tornou-se uma nova área de pesquisa na IA [Freitas, Stuckenschmidt e Noy 2005].

As ontologias ganharam maior "popularidade" quando começaram a ser usadas para domínios mais específicos, ao contrário do que era feito no princípio, quando usadas em SE, onde eram mais gerais e possuíam definições sobre tudo, na tentativa de representação do senso comum.

Segundo Maedche em [Maedche 2002], o principal uso de ontologias na Computação se refere à construção de um artefato, constituído por um vocabulário específico usado para descrever uma certa realidade, mais um conjunto de hipóteses explícitas relativo aos possíveis significados desse vocabulário.

Grande parte da bibliografia sobre ontologias cita a definição apresentada por Gruber em [Gruber 1995], onde aquele autor define uma ontologia como "*uma especificação explícita de uma conceitualização*". Nesta definição:

- conceitualização é um modelo abstrato de um determinado fenômeno do mundo, usualmente restrito ao domínio deste fenômeno;
- a palavra "explícita" indica que os conceitos e relações do modelo abstrato são explícitos em termos e definições.

[Guarino 1996] complementa a definição de ontologia dada por Gruber: "*uma especificação explícita e* ***formal*** *de uma conceitualização* ***compartilhada***". Desta forma, pretende que a ontologia possa ser processada por computador. É necessário representá-la, então, em uma linguagem "formal", o que exclui o uso de uma linguagem natural. Nesta definição, a qualificação "compartilhada" descreve um conhecimento consensual, aceito por um grupo e utilizado por mais de um indivíduo.

Vários conceitos a respeito de ontologias na Ciência da Computação são encontrados na literatura atualmente. Alguns destes, em obras como [Holsapple e Joshi 2002], [Fensel 2002] e [Chandrasekaran, Josephson e Benjamins 1999], chegam a um consenso sobre ontologias: uma ontologia identifica classes - cada uma caracterizada por propriedades que todos os elementos desta classe compartilham - e as organiza hierarquicamente. Isto também inclui importantes relações entre classes e elementos, em um domínio de conhecimento específico. Os autores Jurafsky e Martin, em [Jurafsky e Martin 2000], referem-se a ontologias como um conjunto de objetos distintos resultantes de uma análise de um domínio, ou de um "micromundo".

Provendo estes "recursos" (classes, propriedades, relações, etc) sobre um determinado domínio, uma ontologia permite que aplicações usem uma semântica formal, precisa e clara para processar as informações nela descritas, empregando estas informações em aplicações inteligentes [Freitas, Stuckenschmidt e Noy 2005].

Em nosso trabalho ampliamos o conceito de ontologia para visualizá-la como uma estrutura ontológica, ou seja, um conjunto de termos previamente definidos, associados de forma explícita por meio de relações semânticas, em formato legível por humanos e por máquinas, aí incluindo-se coleções de vocabulários ou de conceitos [Chaves 2003].

Para Maedche em [Maedche 2002], as principais áreas de aplicação de ontologias são a Web Semântica, a Compreensão da Linguagem Natural, a Gestão do Conhecimento e o Comércio Eletrônico.

De acordo com Berners-Lee em [Berners-Lee, Hendler e Lassila 2001], a Web Semântica tem como objetivo dar significado semântico às páginas Web, e tem como apoio ontologias expressas em linguagem formal, as quais definem o conteúdo das páginas Web de forma padronizada, permitindo o compartilhamento de termos e a troca de informações entre páginas Web. Na Gestão do Conhecimento as ontologias são utilizadas como ferramenta para gerenciar os bens intelectuais de uma organização. Permitem a representação estrutural e semântica de documentos, fornecendo novas possibilidades como: busca inteligente, resposta de consultas, troca de documentos, etc. Algumas aplicações que fazem uso de ontologias para gerenciar o conhecimento em uma organização

podem ser conferidas em trabalhos como [Edgington *et al.* 2004], [Houari e Far 2004] e [Everett *et al.* 2002]

Segundo Chandrasekaran e co-autores, a utilidade das ontologias na compreensão da linguagem natural se dá de duas maneiras [Chandrasekaran, Josephson e Benjamins 1999]. Primeiramente provê uma representação do conhecimento de domínio, importante para o processo de desambiguação. Em segundo lugar, ontologias de domínio ajudam na tarefa de identificação de categorias semânticas que estão envolvidas no entendimento do discurso daquele domínio. Alguns exemplos de ontologias com o intuito de compreender a linguagem natural são: *WordNet*[1], *CYC*[2], *GUM*[3]. No Comércio Eletrônico as ontologias têm o papel de fazer a descrição semântica de produtos contidos nas páginas Web, diferentemente da representação da Web atual, que se dá em linguagem natural [Fensel *et al.* 2003] [Maedche 2002]. Esta representação facilitará o entendimento das descrições dos produtos para que agentes de software procurem por exemplo, pelo menor preço de um produto em diferentes páginas Web.

Nesses contextos, uma ontologia descreve tanto o vocabulário de um domínio, quanto o conhecimento representado por um conjunto de termos associados a esse domínio. Em outras palavras, as ontologias são similares a bases de conhecimento que descrevem conceitos por meio de definições, e que são formalmente e suficientemente explicitadas para que haja um entendimento semântico (interoperabilidade) de um determinado domínio [Chandrasekaran, Josephson e Benjamins 1999].

2.2 Mapeamento entre ontologias

O uso e reuso do conhecimento é fundamental, atualmente, dada a impressionante quantidade de informação que está sendo continuamente gerada, forçando organizações e pessoas a gerenciar o conhecimento de modo mais eficaz e eficiente. O fato de combinar conhecimento de domínios distintos acarreta vários problemas como, por exemplo, formatos de representação distintos ou inconsistências semânticas, entre outros [Ding e Foo 2002]. As inconsistências semânticas podem ser geradas quando dois sistemas projetados independentemente são integrados, ocorrendo alguns "mal entendidos" entre os significados e as interpretações para o mesmo dado, tais como nomes, estruturas, *schemas* ou atributos, entre outros [Chaves 2003].

A tarefa de mapeamento de ontologias preocupa-se com o uso e reuso de ontologias, tendo em vista expansão e combinação destas últimas, com o intuito de aumentar a informação e o conhecimento em diferentes domínios que são integrados para suportar nova comunicação e uso [Ding e Foo 2002]. De acordo com Shvaiko e Euzenat em [Shvaiko e Euzenat 2005], o mapeamento é uma operação crítica em diferentes níveis de aplicações como: Web Semântica, integração de *schemas* ou ontologias, Comércio Eletrônico, etc. A operação de mapeamento tem como entrada dois *schemas* ou ontologias, cada um consistindo em um conjunto discreto de entidades (por exemplo, tabelas, elemen-

[1] O acesso *on-line* é feito em: http://wordnet.princeton.edu/

[2] Maiores informações sobre o projeto podem ser encontradas em http://www.cyc.com/

[3] Informações sobre esta ontologia estão em http://purl.org/net/gum2

tos XML, propriedades, regras, predicados), e determina como saída os relacionamentos (equivalência, classificação) existentes entre essas entidades.

Muitos trabalhos encontrados na literatura relacionam em diferentes perspectivas (Inteligência Artificial, Sistemas de Informação, Banco de Dados) as abordagens de mapeamento existentes, como em: [Kalfoglou e Schorlemmer 2003], [Shvaiko e Euzenat 2005], [Noy 2004], [Doan e Halevy 2005]. Embora exista diferença no mapeamento de *schemas* de bases de dados e ontologias, por exemplo, acredita-se que as técnicas desenvolvidas em cada caso possibilitem benefícios para ambas.

Em [Euzenat *et al.* 2004], Euzenat e seus co-autores relatam que, em determinados domínios, é esperado que as ontologias não sejam estáticas, e que várias versões estejam disponíveis. Ontologias interdisciplinares podem ser criadas a partir de ontologias de domínio específico. Sendo assim, diferentes versões de um único domínio serão unificadas e novas informações serão encontradas, com a integração das ontologias existentes. Novas ontologias podem ainda ser construídas pela união de bases de dados heterogêneas ou outras fontes de informação.

Alguns guias para construção de ontologias encontrados na literatura sugerem procedimentos que procuram por ontologias existentes que pertencem ao domínio de interesse e que, eventualmente, possam ser reutilizadas, expandidas ou combinadas, de algum modo, dando origem à nova ontologia. Algumas fontes que apontam para esta alternativa são: [Noy e McGuinness 2001], [Uschold e King 1995], [Gruninger e Lee 2002].

De acordo com Ding e Foo em [Ding e Foo 2002], as ontologias são desenvolvidas para prover semântica comum a agentes de comunicação. Quando dois agentes precisam se comunicar ou trocar informações, é necessário que haja um consenso na comunicação entre os mesmos. Isto conduz à necessidade de mapear os termos usados pelos agentes, entre suas ontologias.

Mapear ontologias construídas individualmente consiste em uma tarefa complexa, a qual traz alguns problemas para sua realização. Os autores Maedche e Staab em [Maedche e Staab 2002] relatam que o inconveniente é que as ontologias reais que conhecemos não especificam suas conceitualizações somente por estruturas lógicas, mas por uma referência a termos, fundamentada em linguagem natural. Ontologias podem também ter más combinações em nível de modelo estrutural como, por exemplo, taxonomias distintas [Davies, Fensel e Harmelen 2003], [Noy 2004].

As ontologias podem se diferenciar quanto à linguagem de representação, o que significa que elas foram representadas com sintaxes diferentes, ou que a expressividade das linguagens usadas para representá-las é diferente. Por exemplo, na construção de uma ontologia em uma determinada linguagem (ver Seção 2.4) define-se que uma classe é parte de outra. Em uma outra linguagem não é possível afirmar a mesma coisa. As semânticas das linguagens são diferentes. Para contornar este problema, ao mapear duas ontologias distintas, o autor Noy [Noy 2004], relata a necessidade de realizar uma normalização, para que as diferenças nas linguagens de representação não interfiram no processo de mapeamento, transformando as ontologias para uma mesma linguagem de descrição.

Atualmente as pesquisas realizadas mostram um grande esforço para automatizar o processo de mapeamento entre ontologias, dado que o mapeamento manual não é escalável

no contexto da Web. De acordo com Noy em [Noy 2004], a construção de ferramentas que permitam o mapeamento automático ou semi-automático entre ontologias deve analisar as seguintes características presentes na definição das ontologias envolvidas no processo:

- *nomes dos conceitos e descrições em linguagem natural*;
- *hierarquia das classes (relacionamento de subclasses-superclasses)*;
- *definições de propriedades (domínio, abrangência, restrições)*;
- *instâncias das classes*;
- *descrições das classes.*

O mapeamento entre ontologias é encontrado em diversas abordagens na literatura. A seguir iremos descrever as duas abordagens identificadas através deste estudo: união e alinhamento.

2.2.1 União

Segundo Sowa *apud* [Pinto, Gómez-Pérez e Martins 1999], o processo de união de duas ou mais ontologias é assim descrito: "É o processo de encontrar associações entre duas ontologias diferentes A e B, e derivar uma nova ontologia C que facilita a interoperabilidade entre sistemas computacionais que são baseados nas ontologias A e B. A nova ontologia C pode substituir A ou B, ou pode ser usada como intermediária entre um sistema baseado em A, e um sistema baseado em B".

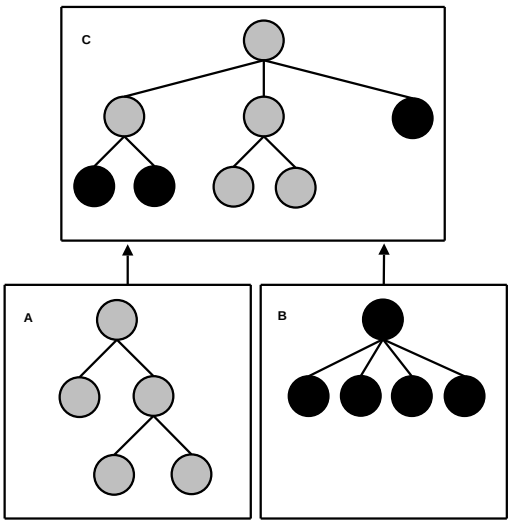


Figura 2.1: União de ontologias

Segundo Chaves em [Chaves 2003], a união entre ontologias consiste na criação de uma nova ontologia a partir de outras já existentes, incluindo na ontologia resultante

as informações provenientes de todas as ontologias envolvidas no processo. A Figura 2.1 apresenta duas ontologias de mesmo domínio que possuem diferenças, tais como termos lexicalmente distintos e estrutura taxonômica diferente, representadas por A e B. Todas estas diferenças são mapeadas e uma nova ontologia C é construída. Esta nova ontologia unifica os conceitos, nomenclaturas, definições e limitações de todas as ontologias envolvidas no processo [Stumme, Studer e Sure 2000].

2.2.2 Alinhamento

Segundo Davies e seus co-autores em [Davies, Fensel e Harmelen 2003], a Web Semântica irá dispor de muitas ontologias de domínio específico com livre acesso. Para formar uma Web "realmente semântica" - o que permitirá que computadores combinem e infiram conhecimento implícito - estas ontologias distintas terão que ser alinhadas e ligadas de algum modo.

De acordo com Sowa *apud* [Pinto, Gómez-Pérez e Martins 1999], o alinhamento faz um mapeamento de conceitos e relações entre duas ontologias A e B que preserva, parcialmente, a ordem dos subtipos em A e B. Se o alinhamento mapeia um conceito ou uma relação x na ontologia A em um conceito ou relação y na ontologia B, então x e y são ditos equivalentes (sinônimos). O alinhamento pode ser parcial, ou seja, podem existir conceitos em A ou B que não possuem equivalentes na outra ontologia. Antes de duas ontologias A e B serem alinhadas, pode ser necessário introduzir novos sub ou supertipos de conceitos ou relações em cada umas das ontologias A e B, para prover "alvos" adequados para o alinhamento. Durante o processo não são necessárias mudanças em axiomas, definições ou provas em A ou B. A Figura 2.2 mostra o processo de alinhamento entre duas ontologias.

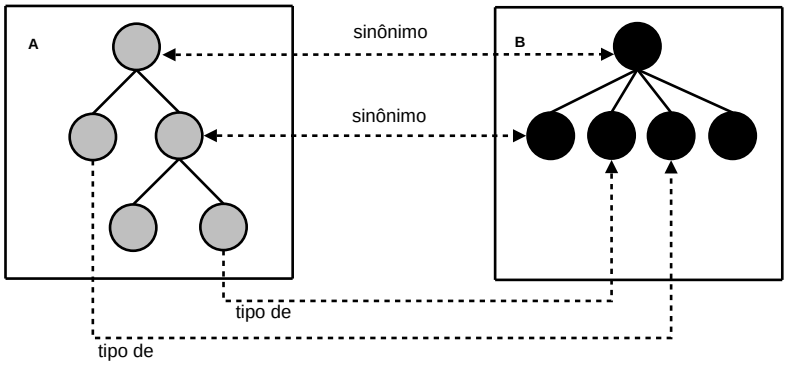


Figura 2.2: Alinhamento entre duas ontologias

O alinhamento de ontologias foi descrito por Euzenat e seus co-autores em [Euzenat et al. 2004] da seguinte maneira: "Dadas duas ontologias que descrevem, cada uma, um conjunto de entidades (classes, propriedades, regras, predicados, etc), o processo de alinhamento consiste em encontrar os relacionamentos (equivalências ou classificações) entre essas entidades".

A interoperabilidade semântica pode ser fundamentada no alinhamento de ontologias: encontrando relacionamentos entre entidades pertencentes a diferentes ontologias. O resultado do alinhamento pode ser usado para vários propósitos, como, por exemplo, mostrar as correspondências, transformar uma fonte em outra ou criar um conjunto de pontes para axiomas ou regras entre ontologias [Euzenat *et al.* 2004].

2.3 Relações semânticas entre termos

As relações entre as palavras podem ser de diversos tipos, e classificam-se em dois grupos: relações semânticas e relações lexicais. As relações semânticas ocorrem entre conceitos, enquanto que as relações lexicais ocorrem entre formas específicas. Muitas vezes, a identificação do tipo de relação entre duas palavras é um tanto dúbia, sendo que é comum encontrarmos, na bibliografia, a expressão "relações semânticas" para identificar tanto relações lexicais quanto relações semânticas.

De acordo com Jurafsky e Martin em [Jurafsky e Martin 2000], as relações semânticas são correspondências existentes entre palavras e seus significados. Jones em [Jones 1986], por sua vez, menciona que o significado de uma palavra pode ser definido por sua relação com outras palavras, e o vocabulário de uma linguagem tem a estrutura determinada por estas relações.

As seções a seguir apresentam os principais tipos de relações semânticas existentes em uma ontologia.

2.3.1 Sinonímia e antonímia

A relação de sinonímia entre duas palavras é dada entre lexemas que possuem o mesmo sentido (sinônimos), ou seja, que possam ser substituídos um pelo outro em uma sentença, sem alterar o significado da mesma [Gasperin 2001]. Por exemplo, a palavra *objetivo* é sinônimo de *propósito*.

Antonímia é a relação de contraste entre duas palavras que possuem sentidos opostos. Duas palavras são consideradas antônimas se exprimem idéias contrárias. Exemplos de palavras antônimas são os pares: *grande* - *pequeno*, *claro* - *escuro*, *novo* - *velho*.

2.3.2 Meronímia e holonímia

A meronímia é a relação semântica entre dois termos, quando um é identificado como parte do outro, ou quando um dos lexemas está contido no outro. Esta relação pode ser entendida como uma relação "**é parte de**". Por exemplo, os lexemas *porta* e *janela* são merônimos ("parte") do lexema *casa*, e *braço* é merônimo ("parte") de *corpo*.

Já a holonímia é a relação inversa à meronímia, isto é, uma relação que ocorre entre dois lexemas quando um contém o outro entre as partes que o constituem. É também entendida com uma relação "**é formado por**" [Yule 1998]. Por exemplo, *casa* é holônimo de *porta* e *janela*, pois pode se dizer que *casa* "é formada por" *porta* e *janela*.

2.3.3 Hiponímia e hiperonímia

De acordo com [Yule 1998], a hiponímia se dá entre dois lexemas, onde o significado de um está contido no significado do outro. Para Gasperin em [Gasperin 2001], pode-se dizer que hiponímia é a relação em que um termo denota ser subclasse do outro. Por exemplo, é o caso do par de palavras *cachorro* - *animal*. A relação entre as palavras dá idéia de especificação. Por exemplo, todo *cachorro* ***é um tipo*** *de animal*. Diz-se, então, que *cachorro* é um hipônimo de *animal*.

A hiperonímia é a relação inversa à hiponímia, isto é, a relação entre dois lexemas, em que um denota uma generalização do outro. Tomemos como exemplo o mesmo par de palavras usado anteriormente para a relação de hiponímia, *cachorro* - *animal*. Temos, então, que *animal* é um hiperônimo (generalização) de *cachorro*.

As relações de hiponímia e hiperonímia podem ser utilizadas para organizar as palavras em uma estrutura hierárquica, onde um lexema é uma generalização de outro, que por sua vez é generalização de outro, e assim por diante. Estas relações podem ser usadas para a construção da taxonomia de uma ontologia.

2.3.4 Homonímia e polissemia

Homonímia é a relação entre duas ou mais palavras que possuem significados distintos, no entanto, possuem a mesma estrutura fonológica. As palavras homônimas podem ser de três tipos:

1. Homógrafas: são as palavras são iguais na escrita e diferentes na pronúncia, por exemplo, *gosto* (substantivo) e *gosto* (verbo na 1ª pessoa do singular do presente do indicativo), bem como, *conserto* (substantivo) e *conserto* (verbo na 1ª pessoa do singular do presente do indicativo);

2. Homófonas: são as palavras iguais na pronúncia e diferentes na escrita. Por exemplo, *cela* (substantivo) e *sela* (verbo); e

3. Homônimos perfeitos: são as palavras iguais na pronúcia e na escrita como, por exemplo, *cura* (verbo) e *cura* (substantivo), *verão* (verbo) e *verão* (substantivo).

A polissemia é a relação entre os diferentes significados de um mesmo lexema que possuem a mesma forma (escrita e pronúncia) [Yule 1998]. Por exemplo, a palavra *banco* (substantivo), pode ter o significado de instituição financeira, ou pode ter o significado de objeto para sentar.

A distinção entre homonímia e polissemia não é muito clara segundo Yule [Yule 1998]. No entanto, uma indicação da diferença entre estas relações são as entradas das palavras em um dicionário. Se uma palavra contém múltiplos significados (polissemia), então terá apenas uma entrada no dicionário, com os seus múltiplos significados enumerados. Se duas palavras são consideradas homônimas, elas terão entradas separadas em um dicionário.

Segundo Chaves em [Chaves 2003], as relações de polissemia e homonímia podem reduzir a precisão em sistemas de recuperação de informação e também no processo de

integração ou mapeamento de informações, pois termos com mesma forma podem ser considerados iguais quando, no entanto, possuem um significado diferente, dependendo do domínio em que estão sendo empregados.

2.4 Linguagens para construção de ontologias

Seguindo a definição dada por Gruber em [Gruber 1993] (descrita na Subseção 2.1), um dos requisitos para uma linguagem usada na construção da ontologia é a sua capacidade de representação formal. Uma linguagem formal permite que as informações descritas sejam processáveis por humanos e por máquinas. Também precisa prover uma definição explícita e não ambígua dos conceitos e relacionamentos descritos na ontologia.

De acordo com Antoniou e co-autores em [Antoniou, Franconi e Harmelen 2005], uma linguagem para construção de ontologias deve especificar, em um nível abstrato (conceitual), o que é necessariamente verdadeiro em um domínio de interesse. Esses autores relatam que uma linguagem para ontologias deve ser capaz de expressar restrições, que declaram o que cada instanciação do domínio pode conter. As linguagens de ontologias devem ainda poder expressar algumas funcionalidades como:

- conceitos (classes) importantes de um domínio;
- relações importantes entre conceitos, que podem ser hierárquicas (relações entre subclasses), outras relações pré-definidas contidas na linguagem, ou definidas pelo usuário (propriedades);
- restrições quanto ao que pode ser expresso (por exemplo, restrições de domínio, abrangência e cardinalidade).

As linguagens de ontologias permitem que os usuários escrevam explicitamente conceitualizações formais de modelos de domínios. Para isto, são necessários alguns requisitos da linguagem como, por exemplo [Antoniou, Franconi e Harmelen 2005]:

- **sintaxe bem definida**: a importância de uma sintaxe bem definida é bem conhecida na área de programação; é uma condição necessária para que a máquina processe a informação nela contida. A maioria das linguagens de ontologias tem a sintaxe baseada em XML;
- **semântica bem definida**: a semântica descreve precisamente o significado do conhecimento, ou seja, a semântica expressa pela linguagem não pode permitir que sejam feitas interpretações subjetivas, e nem que interpretações diferentes sejam possíveis, por diferentes pessoas ou máquinas;
- **suporte eficiente à inferência**: a linguagem deve ser poderosa o suficiente para que possam ser feitas "deduções", através das declarações, usando a notação da linguagem. Estas deduções podem ser feitas através da identificação de membros de uma classe, classificação ou equivalência de classes, entre outras.

Dadas as características "ideais" para uma linguagem de marcação semântica usada na construção de uma ontologia, veremos, nas próximas subseções, alguns exemplos destas linguagens, suas características e sintaxe, bem como suas limitações. Neste trabalho selecionamos duas linguagens para construção de ontologias: RDF e OWL. Ambas são recomendadas pelo *World Wide Web Consortium* (W3C) para a construção de ontologias na Web Semântica.

2.4.1 *Resource Description Framework* - RDF

O *Resource Description Framework* (RDF) é um padrão de metadados para Web desenvolvido pelo W3C. O RDF é uma linguagem de propósito geral para representação de informação na Web, com duas partes principais [Lassila e Swick 2005]:

- RDF: define a maneira como descrever recursos através de propriedades e valores;
- RDF *Schema* (RDFS): define como são formadas as propriedades as quais são utilizadas para definir *schemas*.

O RDF tem por objetivo prover semântica formal às informações, tornando-as processáveis por humanos e por máquinas, facilitando a troca de informações entre aplicativos de forma padronizada, garantindo a interoperabilidade. Este modelo de dados usa XML (*eXtensible Markup Language*) como sintaxe para descrição de seus metadados [Fensel 2000].

Em RDF, um documento contém declarações sobre recursos (pessoas, páginas, etc) que têm propriedades ("é parte de", "é autor de") com certos valores (que podem ser outras pessoas ou páginas). Esta estrutura se torna um modo natural para descrever a maioria dos dados processados por máquinas. Os recursos e valores são identificados, cada um, com um *Uniform Resource Identifier* (URI), assim como nas páginas Web. A estrutura do RDF pode ser representada por sentenças, onde uma propriedade ou valor é associada a um recurso específico. Uma sentença pode ser dividida em três partes: sujeito (recurso), predicado (propriedade do recurso) e objeto (valor da propriedade) [Berners-Lee, Hendler e Lassila 2001].

O RDFS estende o RDF provendo a habilidade de definir o vocabulário do modelo RDF como: classes, propriedades, tipos, domínios, etc. O RDFS oferece um vocabulário distinto para o modelo de classes e propriedades e outras primitivas básicas que podem ser referenciadas pelo modelo RDF [Staab, Erdmann e Maedche 2000].

A seguir, é apresentado um exemplo de sintaxe RDF retirado de [Lassila e Swick 2005], onde é feita a descrição de uma página Web, http://www.ilrt.org/people/cmdjb, que possui três propriedades: title, creator e publisher, sendo seus respectivos valores: Dave Backett's Home Page, Dave Backett e ILRT, University of Bristol. A descrição deste recurso (página Web) é representada em RDF como mostrado na Figura 2.3.

Embora RDF/RDFS seja útil na especificação de ontologias através de conceitos, relações e instâncias, possui algumas limitações e se apresenta como uma linguagem de baixa expressividade semântica, conforme aspectos relacionados em [Fensel 2000]:

- não expressa igualdade e desigualdade entre os dados descritos pelo modelo;
- não aplica cardinalidade e restrições;
- não descreve propriedades tais como: exclusivo, simétrico, transitivo, inverso, ou relações entre propriedades;
- não descreve operações de união, intersecção e complemento;
- o domínio e abrangência só podem ser definidos globalmente.

```
01. <?xml version="1.0"?>
02. <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
03.          xmlns:dc="http://purl.org/dc/elements/1.1/">
04.   <rdf:Description rdf:about="http://www.ilrt.org/people/cmdjb">
05.     <dc:title> Dave Backett's Home Page </dc:title>
06.     <dc:creator> Dave Backett </dc:creator>
07.     <dc:publisher> ILRT, University of Bristol </dc:publisher>
08.   </rdf:Description>
09. </rdf:RDF>
```

Figura 2.3: Exemplo de trecho RDF

Como RDF possui algumas limitações quanto à sua expressividade semântica, é necessário que outras linguagens sejam aplicadas sobre a "camada" RDF, ou seja, linguagens que usem a sintaxe RDF mas adicionem novas classes e propriedades. A seguir será apresentada a linguagem OWL, a qual estende a linguagem RDF.

2.4.2 *Ontology Web Language* - OWL

Conforme [McGuinness e Harmelen 2005], a *Web Ontology Language* (OWL) é uma linguagem para ontologias voltada para a Web. É uma extensão do vocabulário RDF, derivada da linguagem DAML+OIL[4].

Esta linguagem tem como objetivo expressar ontologias de maneira formal, permitindo que seu processamento seja feito tanto por máquinas como por humanos. OWL tem mais facilidades para expressar o significado e a semântica do que XML, RDF e RDFS [Antoniou, Franconi e Harmelen 2005]. A OWL tem como base estas linguagens, bem como suas habilidades de representar conteúdos interpretáveis por máquina na Web.

Conforme [McGuinness e Harmelen 2005], OWL surgiu da necessidade de se ter uma linguagem para especificação de ontologias que fosse compatível com a Web, e não voltada para comunidades específicas. Esta linguagem usa URIs para nomes e RDF para a descrição dos recursos, além de estendê-los adicionando as seguintes capacidades às ontologias:

[4]Esta linguagem para ontologias é uma junção da DAML (*DARPA Agent Markup Language*) com o consórcio europeu o qual desenvolveu a linguagem OIL (*Ontology Inference Layer*). Maiores referências sobre DAML+OIL são encontradas em http://www.w3.org/TR/daml+oil-reference.

- habilidade de ser distribuída através de muitos sistemas;
- escalabilidade para a Web;
- compatibilidade com os padrões da Web para acessibilidade e internacionalização;
- extensibilidade, já que é construída sobre RDF e RDFS e adiciona mais vocabulário para descrever propriedades e classes, entre outros, relações entre classes (disjunções), cardinalidade ("somente um"), igualdade, tipos mais ricos de propriedades, características de propriedades (simetria), e enumerações de classes, entre outras.

A Figura 2.4 apresenta um exemplo da sintaxe OWL, onde é feita a descrição de uma classe denominada *Vintage* (linha 01). Nesta classe é criada uma propriedade chamada *vintageOf* (linha 04) onde é feita uma restrição de cardinalidade (linhas 05 a 07).

```
01. <owl:Class rdf:ID="Vintage">
02.   <rdfs:subClassOf>
03.     <owl:Restriction>
04.       <owl:onProperty rdf:resource="#vintageOf"/>
05.       <owl:minCardinality
06.             rdf:datatype="&xsd,nonNegativeInteger"> 1
07.       </owl:minCardinality>
08.     </owl:Restriction>
09.   </rdfs:subClassOf>
10. </owl:Class>
```

Figura 2.4: Exemplo de trecho OWL

OWL provê três sublinguagens para uso por comunidades específicas de implementação ou usuários [McGuinness e Harmelen 2005]. São elas:

- OWL Lite: dá suporte aos usuários que primeiramente necessitam de uma classificação hierárquica e restrições simples. Provê um caminho para a migração para Tesauro e outras taxonomias. Possui uma complexidade formal menor do que as outras sublinguagens da OWL.
- OWL DL: dedicada aos usuários que necessitam de expressividade máxima para o processamento computacional e decidibilidade (todas as computações serão finalizadas em um tempo finito). OWL DL é assim denominada pois tem correspondência com a lógica de descrições (Description Logic), um campo de pesquisa que estuda a lógica formal em que OWL se baseia.
- OWL Full: é destinada aos usuários que desejam expressividade máxima e liberdade de sintaxe do RDF, sem garantia computacional.

2.5 Considerações sobre este capítulo

Este capítulo apresentou uma fundamentação teórica acerca do assunto maior denominado "ontologias". Foram trazidos os conceitos apresentados pelos principais autores da área, para o termo "ontologia" na Ciência da Computação.

O principal objetivo das ontologias, tal como descrito na bibliografia estudada, é fornecer um entendimento comum e compartilhado a respeito de um determinado domínio. A diversidade de sistemas e aplicações que usam linguagens distintas para a codificação de ontologias torna a comunicação entre estes sistemas uma tarefa muito complexa. Embora existam diferenças de representação, de ontologia para ontologia, podemos identificar algumas características comuns entre as mesmas. Estas características permitem que, entre os elementos de duas ou mais ontologias, se estabeleça algum tipo de correspondência. Tal procedimento é denominado mapeamento.

O mapeamento entre duas ou mais ontologias associa conceitos e relações equivalentes, de diferentes origens, uns com os outros, de acordo com uma relação de similaridade. As abordagens identificadas para o mapeamento de ontologias neste capítulo foram a união e o alinhamento.

Foram descritas, também, as principais relações semânticas existentes em uma ontologia como, por exemplo, as relações de hiponímia e hiperonímia, que são a base da construção de uma estrutura hierárquica, bem como as relações de homonímia e polissemia, que diminuem a precisão das medidas de similaridade. As relações de hiponímia e hiperonímia, bem como relações de superconceitos e subconceitos são analisadas em nossa proposta (ver Capítulo 4) na comparação da similaridade entre ontologias.

Neste capítulo foram ainda descritas linguagens de marcação semântica utilizadas para construção de ontologias. As linguagens mencionadas (RDF e OWL) são recomendações do órgão que padroniza a Web Semântica. Através destas linguagens podemos representar formalmente as relações de hierarquia (hiponímia e hiperonímia), entre outras, existentes em uma estrutura ontológica.

O próximo capítulo apresenta trabalhos voltados ao mapeamento entre ontologias encontrados na literatura e que apresentam técnicas utilizadas para medir a similaridade entre ontologias.

Capítulo 3

Trabalhos correlatos

O mapeamento entre ontologias foi estudado na Seção 2.2. Pudemos identificar que tanto a abordagem de união como a abordagem de alinhamento necessitam de um processo de verificação da similaridade entre os elementos que constituem ontologias ou outras fontes de dados que tenham que ser integradas.

Na literatura estudada, foram encontradas várias estratégias para medir a similaridade entre elementos de ontologias. As mesmas foram classificadas em dois grupos: similaridade lexical e similaridade semântica, embora se observe que esta classificação não distingue exatamente os dois grupos.

A similaridade lexical leva em conta as palavras que constituem os elementos (ou seja, suas cadeias de caracteres). Nesta abordagem normalmente são empregadas soluções que medem a similaridade entre as cadeias de caracteres, e que fazem uso de heurísticas. A medida normalmente é dada através de coeficientes no intervalo [0,1], os quais refletem a proximidade dos elementos, estruturalmente e lexicalmente.

A similaridade semântica visa comparar os elementos de uma ontologia através dos significados dos mesmos, buscando sinônimos ou outras relações semânticas (Seção 2.3) entre estes elementos. A similaridade semântica também é conhecida como similaridade semântico-estrutural. Compara os elementos de acordo com a posição dos mesmos na estrutura hierárquica, buscando as relações semânticas existentes entre estes.

A seguir serão apresentadas abordagens encontradas na bibliografia para medir a similaridade lexical e semântica entre ontologias.

3.1 Similaridade lexical entre ontologias

3.1.1 Similaridade lexical por Combinação de Caracteres

Os autores Maedche e Staab, em [Maedche e Staab 2002], consideram dois níveis para medir a similaridade entre ontologias. São eles o nível lexical, que mede a similaridade através da combinação de caracteres que representam estes termos, e o nível conceitual ou semântico-estrutural (descrito na Seção 3.2.1), que investiga quais relações conceituais podem existir entre os termos. As medidas de similaridade propostas pelos autores

citados, em geral, descrevem a similaridade da especificação de uma ontologia, comparada com outra ontologia.

De acordo com Maedche e Staab em [Maedche e Staab 2002], a comparação feita em nível lexical primeiramente faz uso da Distância de Edição (DE) proposta por Levenshtein [Levenshtein 1966], que se constitui num método para comparar a similaridade entre duas cadeias de caracteres. A DE obtém o número mínimo de operações de edição necessárias para transformar uma cadeia de caracteres em outra. As operações de edição utilizadas para realizar essa transformação são inserção, remoção e substituição de caracteres. Por exemplo, dadas as cadeias de caracteres "computador" e "computadores" a $DE(computador, computadores)$ será igual a 2 pois, para transformar a cadeia de caracteres "computador" na cadeia de caracteres "computadores", são necessárias duas operações de inserção (respectivamente, dos caracteres "e" e "s").

Utilizando a DE, os autores Maedche e Staab em [Maedche e Staab 2002] propõem uma medida de similaridade lexical entre cadeias de caracteres, denominada Combinação de Caracteres (CC), a qual compara duas entradas lexicais L_i e L_j de acordo com a Equação 3.1.

$$CC(L_i, L_j) := \max\left(0, \frac{\min(|L_i|, |L_j|) - DE(L_i, L_j)}{\min(|L_i|, |L_j|)}\right) \in [0, 1] \tag{3.1}$$

A medida CC considera o número de operações de edição que são necessárias para que uma cadeia de caracteres se transforme em outra (este cálculo é feito através da DE), e compara o número destas operações com o comprimento da menor das duas cadeias de caracteres, $(\min(|L_i|, |L_j|))$. A medida CC retorna a similaridade em valores entre 0 e 1, onde 1 denota combinação perfeita e 0 significa ausência de combinação entre as cadeias de caracteres. Tomemos como exemplo as entradas lexicais usadas no exemplo anterior, "computador" e "computadores". Então, o menor comprimento entre os termos é dado por $\min(|computador|, |computadores|)$, ou seja, $min(10, 12)$, portanto 10.

A DE já calculada anteriormente é igual a 2. A subtração $\min(|L_i|, |L_j|)$ - $DE(L_i, L_j)$ tem resultado $10 - 2 = 8$ e a medida CC é dada de acordo com a Equação 3.1 conforme exemplo abaixo.

$$CC(computador, computadores) := \max\left(0, \frac{8}{10}\right) = 0,8 \in [0, 1]$$

Esta medida diminui a influência de pseudo-diferenças nas cadeias de caracteres percebidas em diferentes ontologias como, por exemplo, *underscores*, hífens, singular e plural, e outras marcações adicionais. Um fator importante que deve ser ressaltado é que esta medida pode retornar resultados inadequados, por exemplo, quando duas cadeias de caracteres assemelham-se uma com a outra, mas não existe uma relação de significado entre

as mesmas, como em "power" e "tower" na língua inglesa.

A medida CC consiste na comparação da combinação de caracteres dos termos entre duas ontologias. Como nossa proposta de trabalho busca medir a similaridade semântica, a utilização da CC ocorreria na comparação em nível lexical de dois termos, contribuindo como uma etapa para medir a similaridade semântica. Pesa negativamente o fato de a medida não ser diretamente aplicável a ontologias voltadas para o português.

3.1.2 Proposta de Chaves para a similaridade lexical em ontologias na língua portuguesa

O trabalho desenvolvido por Chaves [Chaves 2003] apresenta uma medida de similaridade lexical para auxiliar no mapeamento semi-automático entre ontologias em língua portuguesa.

O mapeamento entre ontologias inicialmente experimentado com o uso da medida CC (descrita na Seção 3.1.1), não trouxe resultados muito expressivos em nível semântico conforme descrito por Chaves, e termos lexicalmente similares não chegaram a ser mapeados. Chaves relata que a medida CC não é diretamente aplicável a ontologias da língua portuguesa e, para fazê-lo, propõe uma medida denominada "Similaridade Lexical" (SL). A SL usa, em associação à proposta da CC, um algoritmo de *stemming* e uma heurística denominada "Primeira Palavra".

A dificuldade em aplicar diretamente a CC a termos em língua portuguesa se deve às variações que sofrem os termos (exemplo: plural, gerúndio e sufixos de tempo passado). Estas variações, na língua portuguesa, fazem com que a DE usada na medida CC exija um maior número de operações para transformação de uma cadeia de caracteres em outra, produzindo uma combinação ruim do ponto de vista da CC, o que revelaria cadeias de caracteres muito distintas, evitando que termos similares fossem mapeados na ontologia. Foi então empregado um algoritmo de *stemming* para minimizar estes problemas.

De acordo com JurafsKy e Martin em [Jurafsky e Martin 2000], *stemming* é um processo de unir as formas variantes de uma palavra ou palavras lexicalmente similares em uma representação comum, chamada de *stem*. Por exemplo, as palavras *conectar, conectado, conectando* podem ser reduzidas ao *stem* **conect**. O processo de *stemming* é largamente utilizado em processamento de textos para recuperação de informação, no intuito de recuperar documentos que possuam, além da palavra-chave submetida na consulta, as variantes dessa palavra ou palavras lexicalmente similares. Isto é feito através da redução das palavras a seus respectivos *stems*. O algoritmo de *stemming* utilizado por Chaves foi o PortugueseStemmer desenvolvido por Orengo e Huyck [Orengo e Huyck 2001]. Segundo Chaves, este algoritmo foi escolhido para utilização em sua medida devido a melhor precisão obtida em estudo comparativo de *stemmers* para o português do Brasil, de acordo com [Chaves 2003].

A medida SL de Chaves, expressa na Equação 3.2, utiliza um algoritmo de *stemming* associado à medida CC de Maedche e Staab.

$$SL(T_i, T_j) = \min\{\triangle_{ij}^1, \triangle_{ij}^2, ..., \triangle_{ij}^k\} \in [0, 1] \tag{3.2}$$

A medida SL é calculada usando dois termos T_i e T_j, entre os quais se quer medir a similaridade. O termo T_i pertence à ontologia base e o termo T_j pertence à ontologia alvo. Estes termos podem ser monopalavra ou multipalavra. Esta medida leva em consideração somente o *stem* de cada palavra, ao contrário da medida CC, que considera a palavra com todos os seus caracteres. O símbolo $\triangle$ representa o cálculo realizado pela medida CC, de acordo com as seguintes condições:

$$\triangle_{ij}^{k} = \begin{cases} CC(Rad_i^k, Rad_j^k) \text{ se } DE = 0 \\ CC(Rad_i^k, Rad_j^k) - 0.1 \text{ se } DE = 1 \\ CC(Rad_i^k, Rad_j^k) - 0.2 \text{ se } DE = 2 \\ 0 \text{ se } DE \geq 3 \end{cases} \tag{3.3}$$

Os *stems* dos termos T_i e T_j são expressos por Rad_i^k e Rad_j^k, onde o índice k indica a posição da palavra no termo (lembrando aqui que os termos podem ser multipalavra).

Quando os termos T_i e T_j possuem uma quantidade de palavras distintas (exemplo: "Assembléia Legislativa" e "Assembléia Legislativa Estadual"), o índice k varia até a última posição do termo de menor número de palavras. O resultado final da medida SL é o menor valor gerado em $\triangle_{ij}^{k}$, para diferentes k. Esse valor é dependente do valor da DE usada na medida CC.

Na medida SL, a DE é calculada sobre os *stems* e seu resultado é decrementado de acordo com as condições apresentadas na Equação 3.3. Quanto maior for o resultado obtido pela DE, maior será o valor usado para o decremento. Segundo Chaves, os valores 0.1 e 0.2 foram encontrados através de sucessivos experimentos. Quando a DE é maior que 3, então o valor de similaridade retornado pela medida CC em $\triangle_{ij}^{k}$ é igual a zero, pois aquele autor entende que três ou mais alterações no radical de uma palavra caracterizam um baixo grau de similaridade.

Após uma série de testes relatados em [Chaves 2003] para a validação da medida SL, foram detectados alguns casos em que termos não similares estavam sendo considerados similares. Muitos destes termos tinham um alto grau de similaridade lexical, mas semânticas distintas, mesmo ocorrendo que apenas uma letra (por exemplo, a primeira) do par de termos fosse diferente, o que, aliás, é uma das deficiências da medida CC. Considerando este fato, a medida SL foi complementada com uma heurística denominada "Primeira Letra". Esta heurística é apresentada a seguir:

$$\text{Se } Rad[1]_i^k \neq Rad[1]_j^k \text{ então } CC(Rad_i^k, Rad_j^k) = 0$$

Na heurística sugerida por Chaves, o valor 1 entre colchetes significa a posição da primeira letra do radical (ou do *stem*) do termo i ou j. Caso as primeiras letras dos dois radicais (Rad_i^k e Rad_j^k) sejam diferentes, o valor retornado pela medida CC é zero.

A aplicação da medida SL e da heurística "Primeira Letra" são analisadas através de um protótipo desenvolvido por Chaves. Este protótipo faz uma abstração da linguagem usada na construção da ontologia e considera somente a hierarquia da mesma, facilitando o mapeamento e abstraindo a sintaxe. Após o término da execução, é possível ao usuário, através do protótipo, modificar o limiar de similaridade lexical desejado e processar os dados novamente; aceitar ou rejeitar alguns dos mapeamentos; indicar novos mapeamentos não detectados, entre outras funções. Deste modo, Chaves propõe um processo que, em realidade, é de semi-automatização da detecção de similaridade lexical entre termos de ontologias diferentes.

Juntamente com a CC, a medida SL pode contribuir em parte na comparação da similaridade semântica de nosso trabalho, para medir a proximidade lexical de dois termos entre ontologias. O ponto favorável da medida SL é o fato da mesma ser diretamente aplicável em ontologias voltadas ao português, diferentemente da medida CC.

3.1.3 FCA-Merge

O FCA-Merge (*Formal Concept Analysis - Merge*) é um método para unir duas ontologias baseado em Processamento da Linguagem Natural (PLN) e análise formal de conceitos. Este método tem como objetivo derivar um reticulado de conceitos. O resultado é explorado e transformado em uma ontologia unificada, através da interação com um especialista humano. O FCA-Merge foi desenvolvido por Stumme e Maedche e descrito em [Stumme e Maedche 2001].

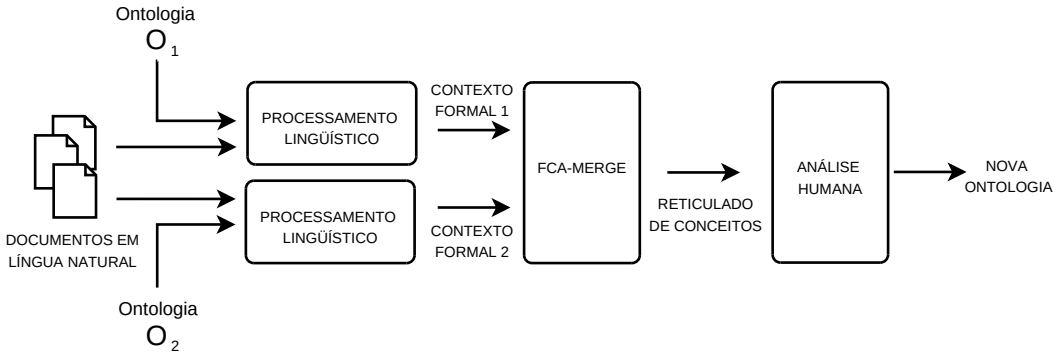


Figura 3.1: Método FCA-Merge (imagem adaptada de [Stumme e Maedche 2001])

Conforme a Figura 3.1, o processo de união de ontologias proposto pelo FCA-Merge tem como entrada duas ontologias O_1 e O_2 e um conjunto de documentos em língua natural - estes documentos devem ser representativos do domínio em questão e devem ter relação com as ontologias. Destes documentos em língua natural são extraídas instâncias relacionadas aos conceitos das ontologias, evitando o problema que ocorre na maioria

das aplicações, onde não existem objetos que são instâncias simultaneamente nas duas ontologias envolvidas no processo de união. As instâncias também são utilizadas como base para identificação de conceitos similares [Kalfoglou e Schorlemmer 2003].

Após, são criados contextos formais (um para cada ontologia), onde é feita a associação dos conceitos das ontologias com palavras ou expressões complexas contidas nos documentos em língua natural. Estes contextos formais são unidos através de técnicas de análise formal de conceitos constituindo um reticulado conceitual. Este reticulado conceitual é então analisado por um especialista humano, tendo como resultado uma nova ontologia.

O método FCA-Merge consiste em três passos, como ilustrado na Figura 3.1. Estes passos são detalhados a seguir:

- **extração de instâncias e criação de contextos formais**: nesta fase são passados como entrada os documentos em língua natural e duas ontologias. São extraídas algumas informações dos documentos através da aplicação de um *tokenizer*, o qual identifica expressões complexas e faz a expansão de abreviaturas de acordo com um léxico pré-definido do domínio. É feita também uma análise lexical nos documentos (análise morfológica; reconhecimento de entidades nomeadas; etiquetagem das partes do discurso). As instâncias são extraídas através da associação de palavras ou expressões complexas com conceitos na ontologia, de acordo com as entradas do léxico específico do domínio. São então construídos os contextos formais para cada ontologia, ou seja, tabelas booleanas indicando qual instância pertence a qual conceito da ontologia.

- **computação de reticulados conceituais**: este segundo passo toma como entrada os contextos formais (um para cada ontologia) gerados no passo anterior. Estes contextos são unidos em um novo contexto formal através de técnicas de análise formal, derivando um reticulado conceitual que será usado como entrada no passo seguinte [Stumme *et al.* 2000].

- **geração interativa da ontologia final unificada**: enquanto os passos anteriores são automáticos, este passo gera a ontologia final através de uma simplificação do reticulado conceitual, feita por um engenheiro de ontologias ou especialista do domínio [Kalfoglou e Schorlemmer 2003].

Uma diferença da abordagem FCA-Merge em relação ao Mapeamento Taxonômico (MT) é a utilização de documentos em língua natural para criar instâncias para as ontologias envolvidas no processo de união, enquanto nossa abordagem emprega apenas a estrutura das ontologias na comparação, não fazendo uso da comparação de instâncias. Um exemplo mais completo do método FCA-Merge para união de ontologias pode ser encontrado em [Stumme e Maedche 2001].

3.2 Similaridade semântica entre ontologias

3.2.1 Mapeamento Taxonômico (MT)

Os autores Maedche e Staab em [Maedche e Staab 2002], descrevem duas abordagens para o mapeamento entre ontologias: a lexical (estudada na Seção 3.1.1) e a conceitual ou semântico-estrutural. A abordagem semântico-estrutural proposta por tais autores compara as hierarquias de duas ontologias, e é denominada Mapeamento Taxonômico (MT).

Para o entendimento das equações e termos com significado específico apresentados a seguir, são trazidas algumas definições e conceitos, de acordo com Cimiano e co-autores [Cimiano, Hotho e Staab 2005].

- **Definição de Ontologia**: uma ontologia é uma estrutura $O := (C, root, \leq_C)$. Onde:
  - C: é um conjunto de identificadores de conceitos;
  - $root$: é o elemento raiz, $root$, representando o elemento no topo da ordem parcial $\leq_C$ em $C \cup \{root\}$, tal que $\forall c \in C \; c \leq root$;
  - $\leq_C$: chamado de hierarquia de conceitos, ou taxonomia.

A abordagem MT utiliza o conceito de "*Semantic Cotopy*" (SC), o qual analisa as relações hierárquicas de subconceito e superconceito do conceito em questão, formando um conjunto de conceitos pertencentes a sua hierarquia [Maedche e Staab 2002], [Dellschaft e Staab 2006], [Cimiano, Hotho e Staab 2004]. O SC é definido por Maeche e Staab em [Maedche e Staab 2002] como o conjunto (C_i) de todos os superconceitos e subconceitos de um conceito c_i em uma ontologia O_i, vide Equação 3.4.

$$SC(c_i, O_i) := \{c_j \in C_i | c_i \leq_C c_j \text{ ou } c_j \leq_C c_i\} \tag{3.4}$$

A partir dos conjuntos de conceitos formados com o SC, é aplicada uma medida que compara a similaridade semântica destes conjuntos, conhecida como **Coeficiente de Jaccard** ou Tanimoto [Manning e Schütze 1999], onde a similaridade de dois conjuntos é dada pela divisão da cardinalidade resultante das operações de intersecção e união destes conjuntos, de acordo com a Equação 3.5.

$$MT(c_i, O_1, O_2) = \frac{|SC(c_i, O_1) \cap SC(c_i, O_2)|}{|SC(c_i, O_1) \cup SC(c_i, O_2)|} \tag{3.5}$$

A Equação 3.5 apresenta uma medida de similaridade que compara semanticamente dois conjuntos de conceitos, formados pelo SC do conceito c_i. No numerador se aplica a cardinalidade da intersecção dos conjuntos de conceitos que foram obtidos através do SC nas duas ontologias (O_1 e O_2), ou seja, o total de elementos comuns em ambas as ontologias. Já no denominador, se aplica a cardinalidade da união dos conjuntos de conceitos que foram obtidos no SC, ou seja, o número total de conceitos que ambas as ontologias (O_1 e O_2) possuem de acordo com o SC.

O resultado final é um coeficiente entre os valores 0 e 1, onde 1 representa uma combinação perfeita dos conceitos, e 0 representa uma má combinação.

Além do conceito do SC descrito por Maedche e Staab em [Maedche e Staab 2002], Cimiano e co-autores em [Cimiano, Hotho e Staab 2005], definem o conceito de "*Common Semantic Cotopy*" (CSC). Ao contrário do SC que forma um conjunto de conceitos contendo todos os superconceitos e subconceitos da hierarquia analisada, o CSC resulta em um conjunto dos superconceitos e subconceitos comuns em ambas as hierarquias examinadas, descartando um conceito que ocorre em apenas uma hierarquia e não ocorre na outra. A definição do CSC é dada pela Equação 3.6.

$$CSC(c_i, O_1, O_2) := \{c_j \in C_1 \cap C_2 | c_i \leq_{C_1} c_j \text{ ou } c_j \leq_{C_1} c_i\} \tag{3.6}$$

Na Equação 3.6, c_i significa o conceito que está sendo analisado. Desta forma, um conceito c_j fará parte do CSC se o mesmo pertence a ambas ontologias O_1 e O_2 ($C_1 \cap C_2$), e se são subconceitos ou superconceitos de c_i, de acordo com a hierarquia de O_1 ($\leq_{C_1}$). Este conjunto de conceitos formado pelo CSC faz com que apenas conceitos comuns às duas ontologias sejam submetidos a comparação da similaridade pela medida de Jaccard. Isto faz com que o coeficiente final de similaridade seja maior (mais próximo ao valor 1), possibilitando encontrar conceitos similares mas em posições hierarquicamente distintas. Também possibilita encontrar más combinações de conceitos como sendo similares, pois conceitos semanticamente distintos podem ter o mesmo CSC.

O MT foi escolhido para ser utilizado, com adaptações, em nosso trabalho, e será melhor explanado no Capítulo 4.

3.2.2 Combinação Semântica

Giunchiglia e Shvaiko em [Giunchiglia e Shvaiko 2004], propõem uma abordagem para a definição da similaridade entre conceitos, denominada Combinação Semântica (CS). O objetivo desta abordagem é explorar os principais problemas de mapeamento entre conceitos similares em estruturas que podem ter sua representação na forma de grafo (por exemplo, hierarquia de conceitos, *schemas* XML, *schemas* de banco de dados ou ontologias).

De acordo com Giunchiglia e Shvaiko, a definição de uma operação para obter a similaridade é dada pela comparação de elementos de duas estruturas de grafos, seguida

da produção de um mapeamento entre os elementos que possuem uma correspondência semântica.

A abordagem CS, proposta por esses autores, apresenta algumas diferenças em relação às abordagens da similaridade lexical:

- a procura pela similaridade é feita através de correspondências semânticas, mapeando significados (conceitos), e não através das palavras que representam os conceitos (rótulos dos nodos), como é feita na similaridade lexical. A CS considera, também, a posição do nodo em relação ao grafo como um todo (por exemplo, análise do nodo irmão, nodos ancestrais, análise de atributos);

- são usadas relações de similaridade semântica entre os conceitos, tais como: igualdade ($=$), intersecção ($\cap$), má combinação ($\perp$), mais geral ($\subseteq$) ou mais específico ($\supseteq$), estes dois últimos são denominados relações de subconjuntos. Ou seja, funciona diferente da similaridade lexical, que geralmente utiliza técnicas orientadas por sintaxe, e medidas de similaridade lexical usando coeficientes no intervalo [0,1], que medem a proximidade lexical entre dois elementos.

A CS calcula a similaridade entre um par de nodos através do significado (conceito) dos rótulos, descritos em língua natural, bem como através do significado estrutural, considerando o par de nodos segundo a posição em que ocupa em relação ao grafo como um todo. É então, atribuída uma das relações semânticas ($=$, $\cap$, $\perp$, $\subseteq$, $\supseteq$) entre o par de nodos, esta relação é transformada em uma fórmula proposicional, que é sumetida a resolvedores denominados SAT (*Boolean Satisfiability*).

A CS traduz formalmente os problemas de comparação de similaridade em grafos, em fórmulas proposicionais, que são submetidas aos sistemas resolvedores SAT , os quais determinam se uma fórmula proposicional é verdadeira ou falsa, ou seja, verificam se a relação semântica atribuída ao par de nodos é válida [Serafini *et al.* 2003].

Giunchiglia e co-autores em [Giunchiglia, Yatskevich e Giunchiglia 2005], descrevem um algoritmo denominado *S-Match* o qual implementa a CS para comparar a similaridade entre duas estruturas representadas por árvore. Este algoritmo computa a relação semântica mais forte contida entre quaisquer pares de nodos. A seguir são apresentados os passos do algoritmo *S-Match*.

1. **Computar conceitos dos rótulos**: como os rótulos dos nodos são representados em língua natural, são utilizados recursos de PLN para extração do conceito que cada nodo representa. Primeiramente são fixados os *tokens* nos rótulos dos nodos. Após, cada *token* é reduzido a sua forma canônica (lema). É então, utilizada a WordNet para capturar o sentido dos lemas dos *tokens*, formando o conceito do rótulo. O conceito de um rótulo é a conjunção de todos os sentidos de cada *token* deste rótulo.

2. **Computar conceitos dos nodos**: o conceito dos nodos é a conjunção do conceito do rótulo obtido no passo anterior, com uma análise da posição do nodo no grafo como um todo. Considera-se o caminho do nodo raiz até o nodo em questão (significado estrutural).

3. **Construir fórmula proposicional**: neste passo são atribuídas as relações semânticas entre os nodos dos grafos, utilizando medidas de similaridade para comparação dos conceitos dos rótulos dos nodos, que servem como entrada para a comparação estrutural dos conceitos dos nodos. Após estabelecida a relação semântica entre o par de nodos, a mesma é traduzida em uma expressão da lógica proposicional (ver Tabela 3.1) para que seja submetida a resolvedores de teoremas.

Tabela 3.1: Regras de transição das relações semânticas para fórmulas proposicionais

rel(a,b)	lógica proposicional
$a = b$	a $\leftrightarrow b$
$a \subseteq b$	$a \rightarrow b$
$a \supseteq b$	$b \rightarrow a$
$a \perp b$	$\neg(a \wedge b)$

   De acordo com a Tabela 3.1, a relação semântica de igualdade (=) é traduzida em equivalência, subconjuntos ($\subseteq$ e $\supseteq$) em implicação e más combinações ($\perp$) em negação da conjunção em lógica proposicional.

4. **Executar SAT**: dada a fórmula proposicional, a mesma é submetida a provadores de teoremas SAT. Ao final os nodos serão mapeados se o resultado retornado pelo SAT for verdadeiro.

Um exemplo mais detalhado da Combinação Semântica pode ser encontrado em [Serafini *et al.* 2003]. A CS utiliza relações semânticas definidas para comparação dos nodos entre duas ontologias, e as transforma em fórmulas proposicionais que são submetidas a resolvedores. Diferente do MT, que utiliza as relações de hierarquia (superconceitos e subconceitos) da própria ontologia formando um conjunto de conceitos (SC) e comparando a similaridade semântica destes conjuntos. Os resolvedores proposicionais da CS não retornam um coeficiente de similaridade entre valores [0,1] como no MT, retornam uma resposta (verdadeiro ou falso) para a consulta, composta de um par de nodos e a relação semântica correspondente.

3.2.3 Cupid

Madhavan e co-autores em [Madhavan, Bernstein e Rahm 2001], apresentam um algoritmo genérico para combinação de *schemas* ou modelos de dados genéricos (XML *schemas*, ontologias, entre outras), envolvendo técnicas lingüísticas e estruturais para comparação da similaridade dos elementos, denominado Cupid. Nesta abordagem, os modelos de dados são transformados em árvores, onde os nodos representam elementos de um modelo de dados específico e são comparados em processo *bottom-up* e *top-down*. O algoritmo de combinação consiste em três fases e opera somente em estruturas de árvore [Euzenat *et al.* 2004].

Similar a algumas abordagens mostradas nas seções anteriores, o algoritmo Cupid computa a similaridade com base em um coeficiente pertencente ao intervalo [0,1], entre

elementos de duas árvores e, então, deduz um mapeamento de acordo com o resultado do coeficiente. A primeira fase, denominada combinação lingüística, computa coeficientes de similaridade lingüística entre os nomes dos rótulos dos nodos. A fase de combinação lingüística é executada em três passos:

1. **Normalização**: normalmente, elementos de *schemas* ou ontologias criadas independentemente possuem nomes que diferem, como por exemplo abreviações, acrônimos, pontuações, etc. É realizado então um passo de normalização. É executada uma *tokenização*, seguida de expansão de abreviaturas e acrônimos e eliminação de *stopwords*. Em cada um desses passos é usado um Tesauro que pode conter tanto termos de uma linguagem comum quanto referências de domínio específico.

2. **Categorização**: os elementos são separados em categorias. Isto é, são agrupados através dos tipos de dados, hierarquia e conteúdo lingüístico. O propósito da categorização, é reduzir o número de comparações de elementos nas duas árvores, considerando somente elementos que pertencem a categorias similares.

3. **Comparação**: o coeficiente de similaridade lingüística (*lsim*) é computado entre os elementos da árvore gerada, pela comparação dos *tokens* extraídos, vide passo anterior. É utilizado o mesmo Tesauro do passo 1, que possui relações de sinonímia e hiperonímia para este propósito. É verificada também a similaridade entre subcadeias de caracteres entre os *tokens*.

O resultado é uma tabela de coeficientes *lsim* entre elementos das duas árvores geradas. O valor *lsim* computado pertence ao intervalo [0,1], com 1 indicando uma combinação lingüística perfeita, e 0 indicando má combinação.

A segunda fase, denominada combinação estrutural, computa um coeficiente de similaridade estrutural entre os elementos de duas árvores, baseado na similaridade de contexto ou vizinhança. Esta fase depende, em parte, da combinação lingüística calculada na primeira fase. O resultado é um coeficiente de similaridade estrutural, denominado de *ssim*, para cada par de elementos. A combinação estrutural é baseada nas seguintes condições:

- elementos atômicos (folhas) em duas árvores são similares se eles são individualmente similares, e se os elementos na sua respectiva vizinhança (ancestrais e irmãos) são similares.

- dois elementos não folhas são similares se eles são lingüisticamente similares, e as raízes das subárvores nos dois elementos são similares.

- dois elementos não folhas são estruturalmente similares se seus conjuntos de folhas são similares, mesmo que seus filhos imediatos não o sejam.

A terceira fase, denominada geração de mapeamento, computa coeficientes de similaridade lingüística e estrutural e gera um mapeamento final escolhendo pares de elementos com pesos de coeficientes de similaridade que são maiores que um limiar previamente

definido [Madhavan, Bernstein e Rahm 2001]. A similaridade final (*wsim*) é dada pela Equação 3.7:

$$wsim = wstruct \times ssim + (1 - wstruct) \times lsim \tag{3.7}$$

onde a constante *wstruct* é um valor entre 0 e 1. O mapeamento é criado pela escolha dos pares dos elementos da árvore com o maior valor de similaridade.

O Cupid, assim como o MT, retorna um coeficiente de similaridade nos valores entre [0,1]. A abordagem Cupid possui duas etapas para a comparação da similaridade: a lingüística e a estrutural. Na comparação lingüística utiliza um Tesauro para normalização dos termos através de *tokenização*, expansão de acrônimos e abreviaturas, e verificação das relações de sinonímia e hiperonímia. A comparação estrutural do MT utiliza o conceito de SC (superconceitos e subconceitos), enquanto o Cupid verifica a similaridade de termos vizinhos, incluindo pais e irmãos dos termos comparados.

3.2.4 Ontology MApping FRAmework - MAFRA

Maeche e co-autores em [Maedche *et al.* 2002], descrevem uma plataforma denominada MAFRA, acrônimo para *Mapping Framework for Distributed Ontologies*, com objetivo de mapear ontologias distribuídas na Web. Este *framework* provê uma visão genérica de todo o processo de mapeamento. Segundo os autores, o processo de mapeamento consiste em um conjunto de atividades para transformar as instâncias de uma ontologia base, em instâncias de uma ontologia alvo.

O *framework* MAFRA consiste de cinco módulos horizontais, os quais descrevem as fases consideradas fundamentais do processo de mapeamento. Também possui quatro módulos verticais que são executados durante todo o processo de mapeamento, em interação com os módulos horizontais. A seguir a descrição dos módulos horizontais de acordo com a Figura 3.2.

Os módulos horizontais são:

1. **Normalização**: este módulo se preocupa em transformar todos os dados que serão mapeados em uma única representação. Lida com a heterogeneidade sintática e estrutural das linguagens utilizadas na construção das ontologias (por exemplo, RDFS, OWL, etc). Elimina as diferenças de sintaxe das linguagens e torna as diferenças semânticas entre as ontologias mais aparentes. Dentro da normalização são utilizadas procedimentos de PLN, como: (*i*) *tokenização* das entidades; (*ii*) eliminação de *stopwords*; (*iii*) expansão de acrônimos.

2. **Similaridade**: este módulo trata a similaridade entre as entidades de uma ontologia base e uma ontologia alvo. São utilizadas diferentes medidas de similaridade para descobrir os mapeamentos. Primeiramente é calculada a similaridade lexical entre cada entidade da ontologia base com todas entidades da ontologia alvo. É utilizada, por exemplo, a WordNet. Também é calculada a similaridade entre as propriedades, pela comparação entre conceitos baseadas nas suas propriedades, atributos ou relações.

3. **Pontes semânticas**: com base nas similaridades computadas na fase anterior, as ligações semânticas são responsáveis por estabelecer correspôndencias entre entidades da ontologia base e da ontologia alvo. Tem como objetivo especificar pontes entre as entidades de maneira que cada instância representada na ontologia base esteja associada a uma instância o mais similar possível, na ontologia alvo.

4. **Execução**: avalia as pontes semânticas construídas entre as entidades das ontologias e transforma instâncias da ontologia base em instâncias na ontologia alvo.

5. **Pós-processamento**: analisa os resultados do módulo de execução com intuito de melhorar a qualidade dos resultados. O principal desafio, segundo os autores, é reconhecer que duas instâncias representam o mesmo objeto do mundo real.

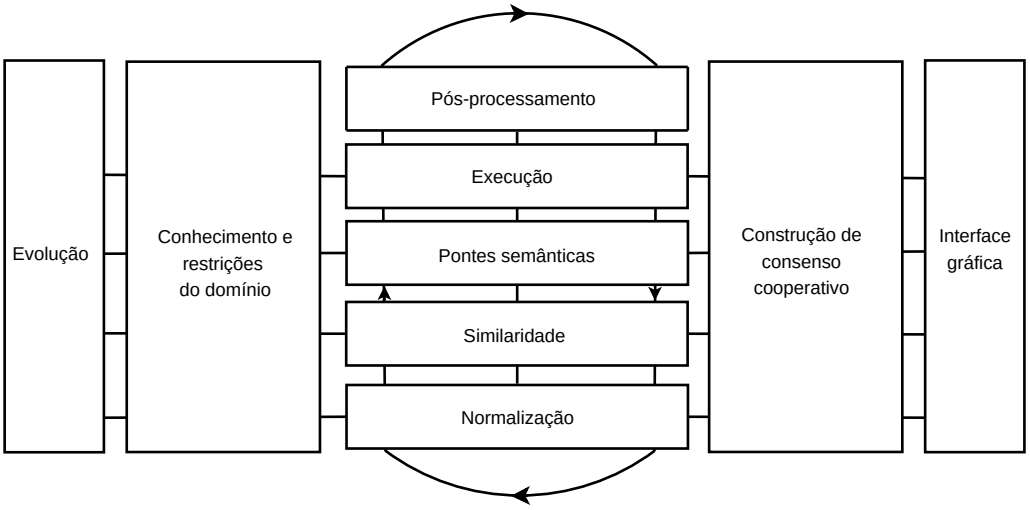


Figura 3.2: Módulos do *framework* MAFRA (adaptado de [Maedche *et al.* 2002])

Os módulos verticais são apresentados a seguir.

1. **Evolução**: este módulo mantém as ligações semânticas construídas no módulo horizontal "Pontes semânticas" à medida que vão ocorrendo mudanças nas ontologias. Este módulo reutiliza as ligações semânticas existentes, fazendo uma adaptação aos novos requisitos do sistema enquanto o mesmo evolui.

2. **Construção de consenso cooperativo**: módulo responsável por estabelecer um consenso das ligações semânticas entre as entidades de duas ontologias, durante o processo de mapeamento entre as entidades. O consenso se faz necessário quando existem muitas alternativas de mapeamento. Este módulo tenta automatizar ao máximo esta tarefa, evitando o envolvimento humano.

3. **Conhecimento e restrições de domínio**: responsável por melhorar a qualidade da similaridade e das pontes semânticas detectadas, introduzindo conhecimento e

restrições de domínio (por exemplo, utilizando glossários para ajudar na identificação de sinônimos, Tesauros e WordNet para identificar conceitos similares).

4. **Interface gráfica do usuário**: permite que o usuário atue como condutor do processo de mapeamento, criando ligações semânticas, refinando-as, etc.

A plataforma MAFRA tem como primeira característica a normalização da sintaxe das ontologias, eliminando as diferenças de sintaxe utilizadas na construção das ontologias. Esta fase de normalização também é adotada em nossa proposta. A MAFRA se diferencia do MT pois a similaridade é computada através da comparação de instâncias das ontologias, bem como de propriedades da mesma. Também utiliza a WordNet para comparação lexical, o que não é feito no MT.

3.2.5 Prompt

Os autores Noy e Musen em [Noy e Musen 2003], desenvolveram um conjunto de ferramentas para manipular ontologias, denominado Prompt. Este *framework* segundo os autores permite ao usuário a utilização de ferramentas para comparar e unir ontologias, manter versões e traduzir para diversos formalismos. O Prompt é uma extensão (*plugin*) do ambiente de edição de ontologias Protegé[1]. Segundo os autores, por ser uma arquitetura de código aberto, o Protegé permite que desenvolvedores possam estender suas funcionalidades através da criação de *plugins* para tarefas específicas.

O Prompt possui duas ferramentas que auxiliam na união semi-automática de ontologias: (i) iPrompt; (ii) ANCHORPrompt.

O iPrompt é uma ferramenta interativa para união de ontologias. Conduz o usuário durante o processo de mapeamento, sugerindo o que pode ser mapeado, identificando inconsistências, possíveis problemas e suas soluções. Esta ferramenta utiliza a estrutura de conceitos em uma ontologia e relações entre os mesmos, bem como a informação que é dada pelo usuário através de suas ações.

O algoritmo do iPrompt tem como entrada duas ontologias, e então, guia o usuário na criação de uma nova ontologia, que é o resultado da união das ontologias dadas como entrada. Após, o algoritmo cria uma lista das combinações feitas através da comparação da similaridade entre os nomes das classes das ontologias. Segundo os autores qualquer medida de similaridade lingüística pode ser utilizada nesta etapa. O algoritmo segue o seguinte ciclo:

1. o usuário dispara uma operação selecionando alguma das sugestões contidas na lista do iPrompt, ou então, dispara uma operação utilizando o ambiente de edição de ontologias para especificar a operação desejada;

2. o iPrompt executa a operação e automaticamente executa algumas mudanças, baseado no tipo da operação, gerando uma lista de sugestões para o usuário baseado na

[1] O Protegé é uma ferramenta de edição de ontologias. As ontologias construídas neste editor podem ser exportadas para uma variedade de formatos (por exemplo, RDFS, OWL, *Schema* XML). Maiores informações no site http://protege.stanford.edu/

estrutura da ontologia e utiliza os argumentos da última operação. Determina inconsistências e possíveis problemas que a última operação pode causar na ontologia e procura sugerir a solução para os mesmos.

O iPrompt permite algumas operações sobre as ontologias envolvidas no processo de mapeamento, são elas: união de classes, instâncias, cópia de uma classe de uma ontologia para outra [Noy e Musen 2003].

A outra ferramenta, o ANCHORPrompt, tem como objetivo determinar possíveis pontos de similaridade entre as ontologias. Esta ferramenta usa a representação da estrutura de grafos das ontologias para encontrar relações entre conceitos e prover informações adicionais ao iPrompt [Noy e Musen 2003].

O algoritmo do ANCHORPrompt tem como entrada duas ontologias e um conjunto de pares-âncoras de termos relacionados, que são identificados através de medidas de similaridade lexical ou definidos pelo usuário. O algoritmo então refina estas combinações analisando o caminho das ontologias de entrada, limitados pelos pares-âncoras, para determinar a freqüência em que os termos aparecem em posições similares ou em caminhos similares. Finalmente, baseado nas freqüências e na resposta do usuário, o algoritmo determina candidatos a combinação [Shvaiko e Euzenat 2005].

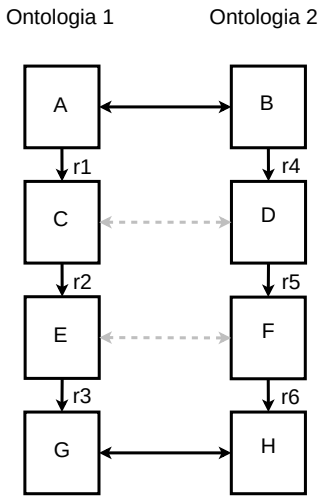


Figura 3.3: Representação do funcionamento do algoritmo ANCHORPrompt

Na Figura 3.3 temos os retângulos representando os termos e os arcos rotulados representando relações semânticas entre os termos de duas ontologias distintas. As setas contínuas entre as ontologias representam os pares-âncoras, e as setas tracejadas indicam pares de termos relacionados.

Como exemplo, consideremos os pares-âncoras na Figura 3.3: A e B, e G e H. Na ontologia 1, composta dos termos A, C, E, G, o comprimento do caminho desde nodo A até o nodo G é 3. Na ontologia 2, composta dos termos B, D, F, H, o comprimento do caminho que vai de B até H também é 3. O grau de similaridade entre os termos C e D e

entre E e F será mais elevado, pois estes termos estão em posições relativas idênticas, no caminho que vai de A até G e de B até H. Sendo assim, o resultado da similaridade entre os termos é cumulativo, ou seja, o grau aumenta à medida que vão sendo encontrados termos similares no caminho.

Segundo Noy e Musen em [Noy e Musen 2003], os algoritmos para mapeamento de ontologias disponíveis no PROMPT possuem forte limitação quando as ontologias possuem hierarquias distintas (por exemplo, profundidades muito diferentes).

O *framework* PROMPT é uma abordagem para união de ontologias o qual analisa a estrutura de um grafo. Diferentemente da abordagem MT que compara um conjunto de termos formados pelas relações hierárquicas dos termos (SC), o PROMPT utiliza pares-ancôras para auxiliar comparação da similaridade entre duas ontologias. Estes pares-ancôras são definidos através de medidas de similaridade lexicais ou através da indicação do usuário pela interface gráfica do *framework*.

3.3 Considerações sobre este capítulo

Neste capítulo foram descritas as principais abordagens utilizadas na literatura para medir a similaridade entre ontologias distintas. Estas abordagens foram classificadas em: similaridade lexical e similaridade semântica.

As medidas de similaridade lexical apresentadas, em sua maioria, fazem uso de um coeficiente de similaridade que varia de 0 a 1, onde 1 denota combinação perfeita dos termos das ontologias, como nas medidas **CC** e **SL**. As abordagens lexicais verificam a similaridade das palavras que representam os termos das ontologias, através da combinação de caracteres que constituem estes termos. O método **FCA-Merge** utiliza análise de conceitos formais e análise lexical para auxiliar um especialista humano a unir duas ontologias. As medidas lexicais conseguem bons resultados quando duas ontologias são lexicalmente similares, ou seja, quando os termos utilizados para descrever os conceitos são escritos com cadeias de caracteres iguais ou muito similares, sendo que o uso de um sinônimo pode não permitir o mapeamento de dois ou mais termos, usando este conceito de similaridade.

As medidas de similaridade lexical não foram o foco maior do estudo bibliográfico, dado que a similaridade lexical possui alguns problemas na identificação da similaridade de alguns conceitos com representações lexicais distintas mas semanticamente similares. Nosso estudo busca alcançar melhores mapeamentos através de medidas de similaridade semântica. Como este trabalho é voltado para ontologias em língua portuguesa vale ressaltar que a medida **SL** foi desenvolvida para ser diretamente aplicada a esta língua.

A similaridade semântica busca comparar os elementos de duas ontologias através da comparação dos significados dos mesmos, buscando sinônimos ou outras relações semânticas entre estes elementos. Algumas medidas de similaridade semântica visam a comparação da estrutura hierárquica das ontologias, como no **MT**, **Cupid** e **Prompt**. Outras medem a similaridade através das relações semânticas que os termos possuem dentro da ontologia e utilizam técnicas de PLN para capturar a semântica como, por exemplo, uso de Tesauros, técnicas de recuperação de informação, bases de dados lexi-

cais (WordNet), como na **CS**, **Cupid** e **MAFRA**.

Com o estudo das principais propostas de medidas de similaridade semântica encontradas na bibliografia, pudemos observar que a maioria utiliza recursos de PLN não diretamente aplicáveis à língua portuguesa como, por exemplo, a WordNet (ainda não disponível para o português), para a identificação de sinônimos. Destacamos o **MT** [Maedche e Staab 2002], que independe do idioma das ontologias, pois não emprega recursos externos de PLN para medir a similaridade. O MT foi escolhido, em nosso trabalho para ser adaptado. As adaptações e primeiros experimentos estão descritos no próximo capítulo.

Capítulo 4

Similaridade Semântica entre ontologias em português

4.1 Medida de *Si*milaridade *Se*mântica - SiSe

4.1.1 Visão Geral

Através do estudo realizado sobre medidas de similaridade lexical e semântica, descritas no Capítulo 3, foi possível analisarmos as características de cada uma destas abordagens, evidenciando seus pontos positivos e negativos. Como nossa proposta visa uma medida de similaridade semântica para ontologias em português, alguns dos trabalhos correlatos não podem ser diretamente aplicados a nossa língua, devido à carência de recursos de Processamento da Linguagem Natural (por exemplo, WordNet, Tesauro de sinônimos, etc) para o português. Desta forma, a abordagem mais natural a ser adaptada é o **Mapeamento Taxonômico** (MT), de Maedche e Staab [Maedche e Staab 2002], descrito na Seção 3.2.1. Em nosso entendimento a utilização de algum dos recursos de PLN disponíveis para o português pode obter ganhos em relação a medida original MT, como será apresentado no decorrer deste capítulo.

O MT, como descrito anteriormente, é uma abordagem que utiliza a similaridade semântico-estrutural entre duas ontologias, ou seja, compara dois termos de ontologias distintas analisando as relações hierárquicas dos mesmos. O MT tem como entrada um termo ou conceito (c_i) pertencente ao léxico formado por termos das duas ontologias, e a estrutura hierárquica das duas ontologias, O_1 e O_2 (conforme Equação 3.5). É feita a comparação deste termo da primeira ontologia dada como entrada, com um termo na segunda ontologia e são comparadas suas hierarquias, através do "*Semantic Cotopy*" ou "*Common Semantic Cotopy*" (conforme as Equações 3.4 e 3.6 respectivamente), os quais analisam os subconceitos e superconceitos de um termo na ontologia.

A medida MT adota uma abordagem que compara a hierarquia das ontologias. Ontologias criadas por diferentes especialistas podem diferir na representação hierárquica para um mesmo conceito, ou seja, cada especialista tem uma visão diferente de um determinado domínio, e estas diferenças são visíveis através da construção de hierarquias

distintas. Este fato, faz com a medida MT que é baseada nos superconceitos e subconceitos dos termos, possa mascarar ou ressaltar algumas similaridades entre os termos das ontologias. Termos que são semanticamente similares podem estar dispostos na hierarquia de tal forma que seus superconceitos e subconceitos sejam diferentes, fazendo com que a medida MT retorne um coeficiente de similaridade baixo. No entanto, esta similaridade hierárquica pode ressaltar similaridades como termos com representações léxicas distintas mas que, no entanto, possuem superconceitos e subconceitos similares, o que pode indicar que estes termos são similares semanticamente.

A título de exemplo, a Tabela H apresenta a estrutura hierárquica de duas ontologias que serão comparadas com a abordagem MT utilizando SC e CSC. Primeiramente, passemos à comparação das ontologias utilizando o SC. As Tabelas 4.2 e 4.3 apresentam o SC dos termos nas ontologias O_1 e O_2, respectivamente.

Tabela 4.1: Trechos de hierarquias do domínio do direito extraídas de duas ontologias

Ontologia 1 (O_1)	Ontologia 2 (O_2)
01 direito constitucional	01 direito
02 direito eleitoral	02 direito eleitoral
03 campanha eleitoral	03 crime eleitoral
04 eleição	04 domicílio eleitoral
05 partido político	05 eleições
06 sistema eleitoral	06 justiça eleitoral
07 voto	07 partidos políticos
	08 sistema distrital
	09 voto

Tabela 4.2: SC dos termos de O_1

Conceitos (O_1)	$SC(C, O_1)$
direito constitucional	{direito constitucional, direito eleitoral, campanha eleitoral, eleição, partido político, sistema eleitoral, voto}
direito eleitoral	{direito constitucional, direito eleitoral, campanha eleitoral, eleição, partido político, sistema eleitoral, voto}
campanha eleitoral	{direito constitucional, direito eleitoral, campanha eleitoral}
eleição	{direito constitucional, direito eleitoral, eleição}
partido político	{direito constitucional, direito eleitoral, partido político}
sistema eleitoral	{direito constitucional, direito eleitoral, sistema eleitoral}
voto	{direito constitucional, direito eleitoral, voto}

Tabela 4.3: SC dos termos de O_2

Conceitos (O_2)	$SC(C, O_2)$
direito	{direito, direito eleitoral, crime eleitoral, domicílio eleitoral, eleições, justiça eleitoral, partidos políticos, sistema distrital, voto}
direito eleitoral	{direito, direito eleitoral, crime eleitoral, domicílio eleitoral, eleições, justiça eleitoral, partidos políticos, sistema distrital, voto}
crime eleitoral	{direito, direito eleitoral, crime eleitoral}
domicílio eleitoral	{direito, direito eleitoral, domicílio eleitoral}
eleições	{direito, direito eleitoral, eleições}
justiça eleitoral	{direito, direito eleitoral, justiça eleitoral}
partidos políticos	{direito, direito eleitoral, partidos políticos}
sistema distrital	{direito, direito eleitoral, sistema distrital}
voto	{direito, direito eleitoral, voto}

Considerando as hierarquias das ontologias na Tabela H, comparamos o termo eleição em O_1, com o termo eleições em O_2. Para cada ontologia consideramos os superconceitos e subconceitos do termo em questão, de acordo com o SC. Na ontologia O_1 os superconceitos do termo eleição são direito eleitoral e direito constitucional e este termo não possui subconceitos. Na ontologia O_2 o termo eleições também não possui subconceitos, e tem direito eleitoral e direito como superconceitos.

O SC cria um conjunto com o termo inicial e seus subconceitos e superconceitos. Temos, então, para o termo eleição em O_1, o conjunto {direito constitucional, direito eleitoral, eleição} e, para o termo eleições em O_2, o conjunto {direito, direito eleitoral, eleições}. Sobre esses dois conjuntos de conceitos é aplicada como medida de similaridade semântica a medida de Jaccard (ver Seção 3.2.1), levando à seqüência de resultados mostrada a seguir, donde se obtém, como coeficiente de similaridade semântica, o valor 1/5.

$$MT = \frac{|SC(\text{eleição}, O_1) \cap SC(\text{eleições}, O_2)|}{|SC(\text{eleição}, O_1) \cup SC(\text{eleições}, O_2)|}$$

$$MT = \frac{|\{\text{direito constitucional}, \text{direito eleitoral}, \text{eleição}\} \cap \{\text{direito}, \text{direito eleitoral}, \text{eleições}\}|}{|\{\text{direito constitucional}, \text{direito eleitoral}, \text{eleição}\} \cup \{\text{direito}, \text{direito eleitoral}, \text{eleições}\}|}$$

$$MT = \frac{|\{\text{direito eleitoral}\}|}{|\{\text{direito}, \text{direito constitucional}, \text{direito eleitoral}, \text{eleição}, \text{eleições}\}|}$$

$$MT = \frac{1}{5} \in [0, 1]$$

O resultado do coeficiente de similaridade fica entre os valores 0 e 1, onde 1 representa uma combinação perfeita, e 0 representa uma má combinação dos termos, de acordo com

as hierarquias. A Tabela 4.4 apresenta os coeficientes de similaridade comparando os termos de O_1 com os termos de O_2, de acordo com o MT, utilizando o SC.

Tabela 4.4: Resultados do experimento (termos de O_1 comparados aos termos de O_2) de acordo com o MT utilizando SC

$O_1 \setminus O_2$	1	2	3	4	5	6	7	8	9
1	0.14	0.14	0.11	0.11	0.11	0.11	0.11	0.11	0.25
2	0.14	**0.14**	0.14	0.11	0.11	0.11	0.11	0.11	0.25
3	0.09	0.09	0.2	0.2	0.2	0.2	0.2	0.2	0.2
4	0.09	0.09	0.2	0.2	**0.2**	0.2	0.2	0.2	0.2
5	0.09	0.09	0.2	0.2	0.2	0.2	**0.2**	0.2	0.2
6	0.09	0.09	0.2	0.2	0.2	0.2	0.2	0.2	0.2
7	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	**0.5**

Através da análise por um humano, temos a identificação de quatro mapeamentos entre os termos das ontologias O_1 e O_2. Os termos considerados similares por análise manual estão descritos na Tabela 4.5.

Tabela 4.5: Termos dados como similares por análise humana

Conceito em O_1	"é similar a"	Conceito em O_2
direito eleitoral	$\Longleftrightarrow$	direito eleitoral
eleição	$\Longleftrightarrow$	eleições
partido político	$\Longleftrightarrow$	partidos políticos
voto	$\Longleftrightarrow$	voto

A partir da identificação manual dos termos similares detectados entre as ontologias, apresentada na Tabela 4.5, temos como comparar os resultados obtidos com a medida MT utilizando o SC, de acordo com a Tabela 4.4. Observamos que os termos considerados similares por humanos tiveram um coeficiente de similaridade relativamente baixo na aplicação da medida MT com o SC (valores em negrito na Tabela 4.4).

Os termos direito eleitoral em O_1 e direito eleitoral em O_2 tiveram coeficiente de similaridade 0.14, o que não retrata a realidade, pois são semanticamente similares. Isto ocorreu pelo fato de seus superconceitos serem distintos (direito constitucional em O_1 e direito em O_2) e, também, pelo fato de direito eleitoral em O_2 possuir um maior número de subconceitos que O_1, diminuindo o coeficiente de similaridade entre os termos.

Os pares de termos eleição em O_1 e eleições em O_2 e partido político em O_1 e partidos políticos em O_2, possuem ambos um coeficiente de similaridade de valor 0.2. Este baixo coeficiente se deve à diferença entre seus superconceitos, e também ao fato de constituírem variações lexicais (em ambos os pares, os termos de O_1 estão no singular e os de O_2 estão no plural), tornando os conjuntos formados pelo SC muito distintos.

O maior coeficiente de similaridade detectado pelo SC do MT, foi aquele entre os termos voto em O_1 e voto em O_2, com um coeficiente de 0.5. Este valor pode ser explicado

pelo fato de os dois termos serem lexicalmente idênticos e, no entanto, possuírem um de seus superconceitos distintos (direito constitucional em O_1 e direito em O_2). Este coeficiente ainda é considerado baixo, pois o valor 0.5 está bem distante do valor que representa combinação perfeita entre os termos, considerando que as hierarquias onde se encontram estes termos são muito similares, bem como o fato de os termos serem idênticos lexicalmente. Apesar de idênticos lexicalmente, os termos tem suas hierarquias analisadas pelo SC, retornando a similaridade semântico-estrutural entre os mesmos.

Pudemos observar que a utilização do SC resulta um coeficiente baixo para termos semanticamente similares. Temos, então, que tratar questões como termos com maior quantidade de subconceitos ou superconceitos em uma ontologia em relação a um termo em outra ontologia (taxonomias distintas), bem como as pequenas diferenças lexicais em que os termos podem estar representados (representações distintas em linguagem natural, exemplo: plural).

Comparamos, também, as ontologias apresentadas na Tabela H utilizando o CSC da medida MT, o qual foi descrito na Seção 3.2.1. As Tabelas 4.6 e 4.7 apresentam o CSC de cada termo das ontologias O_1 e O_2 respectivamente.

O CSC define, para um termo de uma ontologia, um conjunto formado pelo próprio termo, e pelos subconceitos e superconceitos desse termo que são comuns em ambas as ontologias ($C_1 \cap C_2$). Para as ontologias que estamos comparando (Tabela H), os conceitos comuns são: {direito eleitoral, voto}. Os "conceitos comuns em ambas as ontologias" se referem a conceitos cuja forma é lexicalmente idêntica em O_1 e O_2.

Tabela 4.6: CSC dos termos de O_1

Conceitos (O_1)	$CSC(C, O_1, O_2)$
direito constitucional	{direito constitucional, direito eleitoral, voto}
direito eleitoral	{direito eleitoral, voto}
campanha eleitoral	{direito eleitoral, campanha eleitoral}
eleição	{direito eleitoral, eleição}
partido político	{direito eleitoral, partido político}
sistema eleitoral	{direito eleitoral, sistema eleitoral}
voto	{direito eleitoral, voto}

A utilização do CSC faz com que o problema ocorrido na comparação dos conjuntos formados pelo SC (um termo em uma ontologia possuindo um número maior de superconceitos ou subconceitos do que em outra ontologia) baixando o coeficiente final de similaridade, seja sanado, como veremos a seguir.

Quando comparamos o conceito direito eleitoral em O_1 e direito eleitoral O_2, utilizando o SC temos como coeficiente de similaridade o valor 0.14 (vide Tabela 4.4). Estes termos apesar de lexicalmente e semanticamente similares, apresentam diferenças em suas hierarquias. Por exemplo, em O_1 o termo direito eleitoral possui sete elementos, e em O_2 o mesmo termo possui nove elementos, no conjunto formado pelo SC. Esta diferença no número de elementos dos conjuntos reflete algumas diferenças entre as hierarquias das ontologias e, ao aplicarmos a medida de Jaccard, teremos uma baixa no coeficiente de

similaridade entre termos semanticamente similares.

Tabela 4.7: CSC dos termos de O_2

Conceitos (O_2)	$CSC(C, O_2, O_1)$
direito	{direito, direito eleitoral, voto}
direito eleitoral	{direito eleitoral, voto}
crime eleitoral	{direito eleitoral, crime eleitoral}
domicílio eleitoral	{direito eleitoral, domicílio eleitoral}
eleições	{direito eleitoral, eleições}
justiça eleitoral	{direito eleitoral, justiça eleitoral}
partidos políticos	{direito eleitoral, partidos políticos}
sistema distrital	{direito eleitoral, sistema distrital}
voto	{direito eleitoral, voto}

Ao compararmos os termos direito eleitoral em O_1 e direito eleitoral em O_2 utilizando o CSC, temos conjuntos formados somente pelos termos comuns a ambas as ontologias. Por exemplo, os termos comuns a O_1 e O_2 são os seguintes: {direito eleitoral, voto}. Desta forma o CSC analisa a hierarquia do termo direito eleitoral em O_1, o qual possui o superconceito direito constitucional e os subconceitos campanha eleitoral, eleição, partido político, sistema eleitoral e voto. No entanto somente os superconceitos e subconceitos comuns às duas ontologias formarão este conjunto. Desta forma o conjunto formado pelo CSC para este termo em O_1 é {direito eleitoral, voto}. Funcionando da mesma maneira para o termo direito eleitoral em O_2, que possui o conjunto {direito eleitoral, voto} formado pelo CSC.

Ao aplicarmos a medida de Jaccard a estes dois conjuntos, obtemos um coeficiente de similaridade de valor 1, ou seja, uma combinação perfeita entre os termos comparados. A seguir temos o exemplo deste cálculo utilizando o conceito de CSC.

$$MT = \frac{|CSC(\text{direito eleitoral}, O_1, O_2) \cap CSC(\text{direito eleitoral}, O_2, O_1)|}{|CSC(\text{direito eleitoral}, O_1, O_2) \cup CSC(\text{direito eleitoral}, O_2, O_1)|}$$

$$MT = \frac{|\{\text{direito eleitoral}, \text{voto}\} \cap \{\text{direito eleitoral}, \text{voto}\}|}{|\{\text{direito eleitoral}, \text{voto}\} \cup \{\text{direito eleitoral}, \text{voto}\}|}$$

$$MT = \frac{|\{\text{direito eleitoral}, \text{voto}\}|}{|\{\text{direito eleitoral}, \text{voto}\}|}$$

$$MT = \frac{2}{2} \in [0, 1]$$

A Tabela 4.8 apresenta os coeficientes de similaridade dados pelo MT utilizando CSC, entre os termos das ontologias O_1 e O_2.

De acordo com os resultados apresentados na Tabela 4.8 pudemos observar que, para alguns termos não similares, os coeficientes foram altos, não retratando a realidade. Por exemplo, os termos direito eleitoral (O_1) e voto (O_2) resultaram em um coeficiente de valor 1 pois, de acordo com o CSC, possuem o mesmo conjunto de termos conforme suas hierarquias ({direito eleitoral, voto}, ver Tabela 4.6 e 4.7), o que não retrata a realidade.

Tabela 4.8: Resultados do experimento (termos de O_1 comparados aos termos de O_2) de acordo com o MT utilizando CSC

$O_1 \setminus O_2$	1	2	3	4	5	6	7	8	9
1	0.5	0.66	0.25	0.25	0.25	0.25	0.25	0.25	0.66
2	0.66	**1.0**	0.33	0.33	0.33	0.33	0.33	0.33	1.0
3	0.25	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
4	0.25	0.33	0.33	0.33	**0.33**	0.33	0.33	0.33	0.33
5	0.25	0.33	0.33	0.33	0.33	0.33	**0.33**	0.33	0.33
6	0.25	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.33
7	0.66	1.0	0.33	0.33	0.33	0.33	0.33	0.33	**1.0**

A utilização do CSC produz um aumento no coeficiente de similaridade quando um termo possui superconceitos ou subconceitos distintos e que não são comuns às duas ontologias, como ao comparar voto em O_1 e voto em O_2. Estes dois termos possuem superconceitos distintos (direito constitucional em O_1 e direito em O_2). Como o CSC inclui somente conceitos que são similares em ambas as ontologias, estes dois superconceitos não farão parte do conjunto formado pelo CSC, fazendo com que o valor do coeficiente de similaridade semântica entre estes termos aumente.

Ao compararmos os coeficientes de similaridade obtidos com o MT utilizando o CSC e com o SC, apresentados na Tabela 4.9, podemos observar um aumento no coeficiente de similaridade entre os termos considerados similares (manualmente, vide Tabela 4.5) por humanos entre as ontologias. No entanto as diferenças lexicais interferem no coeficiente como, por exemplo, ao compararmos os termos eleição em O_1 e eleições em O_2 e partido político em O_1 e partidos políticos em O_2.

Tabela 4.9: Comparativo dos coeficientes de similaridade entre os termos das ontologias O_1 e O_2 com a medida MT utilizando SC e CSC

Conceito O_1	Conceito O_2	MT(SC)	MT(CSC)
direito eleitoral	direito eleitoral	0.14	1.0
eleição	eleições	0.2	0.33
partido político	partidos políticos	0.2	0.33
voto	voto	0.5	1.0

4.1.2 Adaptações realizadas em relação ao MT

Na seção anterior, foi feita uma descrição mais aprofundada da medida MT, foram apresentados exemplos de sua utilização com duas abordagens para comparação das hierarquias de duas ontologias, a primeira utilizando o "*Semantic Cotopy*" (SC), e a segunda utilizando o "*Common Semantic Cotopy*" (CSC).

Nossa contribuição para medir a similaridade semântica entre ontologias em português é feita através da abordagem denominada **Similaridade Semântica (SiSe)**. Para definirmos a medida SiSe realizamos alguns experimentos através da adaptação da medida MT, no que se refere ao conceito do SC e do CSC. Estas adaptações utilizam um recurso de PLN denominado algoritmo de *steming*. O uso do *stemming* prioriza o nível lexical no intuito de encontrar termos lexicalmente similares para a comparação semântico-estrutural.

O algoritmo de *stemming* transforma as variações lexicais de uma mesma palavra ou palavras lexicalmente similares em uma representação única, chamada *stem*. Paice em [Paice 1994], relata que em uma determinada língua, uma palavra contém um *stem* que se refere a uma idéia central ou significado, e que certos afixos[1] foram adicionados para modificar o significado ou combinar as palavras para exercer uma função sintática. De acordo com Porter em [Porter 2006], o *stemming* é o processo de redução automática de uma ou mais palavras através da retirada de seus sufixos, ignorando o fato da precisão da origem da palavra (raiz). O *stem* não é necessariamente a raiz lingüística ou radical, mas denota uma forma mínima preferencialmente não ambígua do termo [Chaves 2003].

Associados ao processo de *stemming* podem ocorrer dois tipos de erros, denominados de *overstemming* e *understemming*. O erro de *overstemming* ocorre quando uma cadeia de caracteres removida não faz parte do sufixo mas pertence ao *stem*. Por exemplo, a palavra gramática após ser processada por um certo *stemmer* retorna o *stem* grama, sendo que o *stem* correto seria gramát. Já o erro denominado *understemming* ocorre quando o sufixo não é removido completamente. Por exemplo, a palavra referência é transformada no *stem* referênc, ao invés do *stem* considerado correto refer. Nos casos em que o algoritmo de *stemming* apresenta algum destes erros mencionados, os mesmos não são corrigidos em nosso trabalho. Sendo assim, o resultado da medida SiSe está diretamente ligado ao bom funcionamento deste algoritmo.

O algoritmo de *stemming* utilizado em nosso trabalho foi o PortugueseStemmer [Orengo e Huyck 2001], o mesmo utilizado na medida SL de [Chaves 2003], descrita na Seção 3.1.2.

A medida SiSe utiliza o algoritmo de *stemming* para representar os elementos do conjunto formado pelo SC ou CSC. Desta forma o SC' e o CSC' de um termo de uma ontologia têm o conjunto de subconceitos e superconceitos representados pelos seus *stems*.

Para representarmos os *stems* do termos no conjunto formado pelos SC' e CSC', adotamos uma representação para os termos monopalavra e multipalavra. Para termos monopalavra o *stem* é representado em letras minúsculas. Para os termos multipalavra

[1] Os afixos são elementos mórficos que se agregam a uma raiz ou radical a fim de mudar o sentido de uma palavra. Podem ser prefixos, se antepostos ao radical, ou sufixos, se inseridos ao fim do radical [Bechara 2001].

o *stem* correspondente da primeira palavra é representado em letras minúsculas seguido dos *stems* seguintes com a primeira letra maiúscula. Por exemplo, o termo domicílio eleitoral, possui os *stems* domicili eleitor que são representados como domiciliEleitor.

Primeiramente modificamos a equação original do SC (Equação 3.4), e temos as adaptações SC' e SiSe representadas pelas equações 4.1 e 4.2.

$$SC'(c_i, O_1) := \{\Delta_{c_j} \in C_i | c_i \leq_{C_1} c_j \text{ ou } c_j \leq_{C_1} c_i\} \tag{4.1}$$

$$SiSe(c_1, O_1, c_2, O_2) = \frac{|SC'(c_1, O_1) \cap SC'(c_2, O_2)|}{|SC'(c_1, O_1) \cup SC'(c_2, O_2)|} \in [0, 1] \tag{4.2}$$

Na Equação 4.1, o símbolo Δ representa o *stem* do conceito C_j, possibilitando que as variações lexicais dos termos possam ser consideradas similares através de uma única representação, permitindo que termos similares tenham um coeficiente de similaridade semântica maior, na comparação da hierarquia. Estes conjuntos representados pelos *stems* dos termos são comparados através da medida de Jaccard, assim como na medida MT, de acordo com a Equação 4.2.

As hierarquias das ontologias utilizadas neste experimento estão representadas na Tabela H. O SC' de cada termo das ontologias é representado por seu *stem*. Desta forma os conjuntos são representados de acordo com a Tabela 4.10 para os termos em O_1, e na Tabela 4.11 para os termos em O_2.

Tabela 4.10: SC' dos termos de O_1

Conceitos (O_1)	$SC'(C, O_1)$
direito constitucional	{direitConstituc, direitEleitor, campanhEleitor, ele, partPolitic, sistemEleitor, vot}
direito eleitoral	{direitConstituc, direitEleitor, campanhEleitor, ele, partPolitic, sistemEleitor, vot}
campanha eleitoral	{direitConstituc, direitEleitor, campanhEleitor}
eleição	{direitConstituc, direitEleitor, ele}
partido político	{direitConstituc, direitEleitor, partPolitic}
sistema eleitoral	{direitConstituc, direitEleitor, sistemEleitor}
voto	{direitConstituc, direitEleitor, vot}

Como exemplo, comparamos a similaridade dos termos partido político em O_1 e partidos políticos em O_2 de acordo com SC'. Em O_1 partido político tem como superconceitos os termos direito eleitoral e direito constitucional, e não possui subconceitos. O conjunto formado pelo SC' é representado pelo *stem* de cada um destes termos, partPolitic, direitConstituc e direitEleitor, respectivamente, formando o conjunto {partPolitic, direitConsti-

tuc, direitEleitor}. O termo partidos políticos em O_2 não possui subconceitos e tem como superconceitos os termos direito eleitoral e direito. Desta forma o conjunto resultante, de acordo com o SC' é {partPolitic, direitEleitor, direit}.

Tabela 4.11: SC' dos termos de O_2

Conceitos (O_2)	$SC'(C, O_2)$
direito	{direit, direitEleitor, crimEleitor, domiciliEleitor, ele, justicEleitor, partPolitic, sistemDistrit, vot}
direito eleitoral	{direit, direitEleitor, crimEleitor, domiciliEleitor, ele, justicEleitor, partPolitic, sistemDistrit, vot}
crime eleitoral	{direit, direitEleitor, crimEleitor}
domicílio eleitoral	{direit, direitEleitor, domiciliEleitor}
eleições	{direit, direitEleitor, ele}
justiça eleitoral	{direit, direitEleitor, justicEleitor}
partidos políticos	{direit, direitEleitor, partPolitic}
sistema distrital	{direit, direitEleitor, sistemDistrit}
voto	{direit, direitEleitor, vot}

Através dos conjuntos formados pelo SC' dos termos das ontologias O_1 e O_2, aplicamos a medida de Jaccard para compararmos a similaridade semântica dos mesmos, de acordo com a Equação 4.2. Como coeficiente de similaridade entre os dois termos resultou o valor 0.5. Este valor é considerado baixo, e isto pode ser explicado pelo fato da medida comparar a estrutura hierárquica dos termos, ocorrendo que um termo pode ter um maior número de superconceitos e subconceitos do que o outro termo na outra ontologia, diminuindo a similaridade. A seguir observamos as etapas do cálculo da similaridade entre estes dois termos.

$$SiSe = \frac{|SC'(\text{partido político}, O_1) \cap SC'(\text{partidos políticos}, O_2)|}{|SC'(\text{partido político}, O_1) \cup SC'(\text{partidos políticos}, O_2)|}$$

$$SiSe = \frac{|\{\text{direitConstituc, direitEleitor, partPolitic}\} \cap \{\text{direit, direitEleitor, partPolitic}\}|}{|\{\text{direitConstituc, direitEleitor, partPolitic}\} \cup \{\text{direit, direitEleitor, partPolitic}\}|}$$

$$SiSe = \frac{|\{\text{direitEleitor, partPolitic}\}|}{|\{\text{direit, direitConstituc, direitEleitor, partPolitic}\}|}$$

$$SiSe = \frac{2}{4} \in [0, 1]$$

Os termos partido político (O_1) e partidos políticos (O_2), bem como eleição (O_1) e eleições (O_2), obtiveram um aumento em relação ao SC e CSC, devido à utilização do algoritmo de *stemming* que permite que as diferenças lexicais entre estes termos possam

ser minoradas. Em geral, o SC' obteve coeficientes de similaridade maiores do que o SC. Em relação ao CSC o SC' obteve menores coeficientes nos casos em que um termo possui um maior número de subconceitos ou superconceitos em uma ontologia em relação a outro termo na outra ontologia.

A Tabela 4.12 apresenta os coeficientes de similaridade dados pela medida SiSe utilizando o SC'.

Tabela 4.12: Resultados do experimento (termos de O_1 comparados aos termos de O_2) de acordo com o SiSe utilizando SC'

$O_1 \setminus O_2$	1	2	3	4	5	6	7	8	9
1	0.33	0.33	0.11	0.11	0.25	0.11	0.25	0.11	0.25
2	0.33	0.33	0.11	0.11	0.25	0.11	0.25	0.11	0.25
3	0.09	0.09	0.2	0.2	0.2	0.2	0.2	0.2	0.2
4	0.2	0.2	0.2	0.2	0.5	0.2	0.2	0.2	0.2
5	0.2	0.2	0.2	0.2	0.2	0.2	0.5	0.2	0.2
6	0.09	0.09	0.2	0.2	0.2	0.2	0.2	0.2	0.2
7	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.5

Apresentamos, também, uma adaptação na medida MT no que se refere ao conceito do CSC. Modificamos a definição original do CSC, descrito na Equação 3.6, e apresentamos o CSC' e SiSe, de acordo com as Equações 4.3 e 4.4.

$$CSC'(c_i, O_1, O_2) := \{\Delta_{c_j} \in \Delta_{C_1} \cap \Delta_{C_2} | c_i \leq_{C_1} c_j \text{ ou } c_j \leq_{C_1} c_i\} \tag{4.3}$$

$$SiSe(c_1, O_1, c_2, O_2) = \frac{|CSC'(c_1, O_1, O_2) \cap CSC'(c_2, O_2, O_1)|}{|CSC'(c_1, O_1, O_2) \cup CSC'(c_2, O_2, O_1)|} \in [0, 1] \tag{4.4}$$

Na Equação 4.3, o símbolo Δ representa o *stem* dos termos em questão. O conjunto CSC' de um termo é formado com base nos seus subconceitos e superconceitos que são comuns em ambas as ontologias. Estes termos comuns são representados através de seus *stems* Δ_{C_1} e Δ_{C_2}, formando um conjunto $\Delta_{C_1} \cap \Delta_{C_2}$. Desta forma, o *stem* de um subconceito ou superconceito Δ_{C_j} fará parte do conjunto CSC' se o mesmo aparecer em ambas as hierarquias. Os conjuntos dos termos de cada ontologia formados pelo CSC' são comparados através da medida de Jaccard, conforme a Equação 4.4.

As ontologias utilizadas no experimento estão representadas pelas hierarquias apresentadas na Tabela H. O conjunto de termos comuns para estas ontologias de acordo com $\Delta_{C_1} \cap \Delta_{C_2}$ é {partPolitic, vot, direitEleitor, ele}. Com base nesse conjunto de termos comuns é formado o CSC' dos termos das ontologias O_1 e O_2 como apresentado nas Tabelas 4.13 e 4.14.

Por exemplo, ao compararmos a similaridade entre os termos partido político em O_1 e partidos políticos em O_2, temos como CSC' do termo partido político seus superconceitos

direito eleitoral e direito constitucional, não contendo nenhum subconceito. Através do algoritmo de *stemming* temos os respectivos *stems* para os termos: partPolitic, direitEleitor, direitConstituc. Analisando os termos comuns em ambas as ontologias o termo partido político em O_1 tem como CSC' o conjunto {partPolitic, direitEleitor}, pois direitConstituc $\notin \Delta_{C_1} \cap \Delta_{C_2}$.

Tabela 4.13: CSC' dos termos de O_1

Conceitos (O_1)	$CSC'(C, O_1, O_2)$
direito constitucional	{direitConstituc, direitEleitor, ele, partPolitic, vot}
direito eleitoral	{direitEleitor, ele, partPolitic, vot}
campanha eleitoral	{direitEleitor, campanhEleitor}
eleição	{direitEleitor, ele}
partido político	{direitEleitor, partPolitic}
sistema eleitoral	{direitEleitor, sistemEleitor}
voto	{direitoEleitor, vot}

Tabela 4.14: CSC' dos termos de O_2

Conceitos (O_2)	$CSC'(C, O_2, O_1)$
direito	{direit, direitEleitor, ele, partPolitic, vot}
direito eleitoral	{direitEleitor, ele, partPolitic, vot}
crime eleitoral	{direitEleitor, crimEleitor}
domicílio eleitoral	{direitEleitor, domiciliEleitor}
eleições	{direitEleitor, ele}
justiça eleitoral	{direitEleitor, justicEleitor}
partidos políticos	{direitEleitor, partPolitic}
sistema distrital	{direitEleitor, sistemDistrit}
voto	{direitEleitor, vot}

O termo partidos políticos em O_2 também não possui subconceitos, e tem como superconceitos os termos direito eleitoral e direito. Os termos são representados pelos *stems* partPolitic, direitEleitor e direit. Verifica-se quais destes termos são comuns às duas ontologias e tem-se o conjunto {partPolitic, direitEleitor}, formado pelo CSC'. O termo direit não faz parte do conjunto pois não é um termo comum às ontologias.

Dados os dois conjuntos para cada termo, de acordo com o CSC', se aplica a medida de Jaccard para medir a similaridade semântica (Equação 4.4). A seguir apresentamos o exemplo do cálculo de similaridade entre estes dois termos.

$$SiSe = \frac{|CSC'(\text{partido político}, O_1, O_2) \cap CSC'(\text{partidos políticos}, O_2, O_1)|}{|CSC'(\text{partido político}, O_1, O_2) \cup CSC'(\text{partidos políticos}, O_2, O_1)|}$$

$$SiSe = \frac{|\{\text{direitEleitor, partPolitic}\} \cap \{\text{direitEleitor, partPolitic}\}|}{|\{\text{direitEleitor, partPolitic}\} \cup \{\text{direitEleitor, partPolitic}\}|}$$

$$SiSe = \frac{|\{\text{direitEleitor, partPolitic}\}|}{|\{\text{direitEleitor, partPolitic}\}|}$$

$$SiSe = \frac{2}{2} \in [0, 1]$$

A Tabela 4.15 apresenta os resultados obtidos com a medida SiSe utilizando o CSC'. Em relação aos termos dados como similares por análise humana apresentados na Tabela 4.5, o CSC' resultou coeficientes com valores maiores em relação ao SC'. Também verificamos que todos estes termos em questão resultaram coeficiente de similaridade 1, representando uma combinação perfeita dos termos comparados.

Tabela 4.15: Resultados do experimento (termos de O_1 comparados aos termos de O_2) de acordo com o SiSe utilizando CSC'

$O_1 \setminus O_2$	1	2	3	4	5	6	7	8	9
1	0.66	0.8	0.16	0.16	0.4	0.16	0.4	0.16	0.4
2	0.8	1.0	0.2	0.2	0.5	0.2	0.5	0.2	0.25
3	0.16	0.2	0.33	0.33	0.33	0.33	0.33	0.33	0.33
4	0.4	0.5	0.33	0.33	1.0	0.33	0.33	0.33	0.33
5	0.4	0.5	0.33	0.33	0.33	0.33	1.0	0.33	0.33
6	0.16	0.2	0.33	0.33	0.33	0.33	0.33	0.33	0.33
7	0.4	0.5	0.33	0.33	0.33	0.33	0.33	0.33	1.0

Destacamos, o alto coeficiente de similaridade resultado entre alguns termos comparados entre as ontologias, de acordo com a Tabela 4.16. Estes termos, apesar de serem lexicalmente distintos, possuem hierarquias muito similares, ou seja, possuem superconceitos e subconceitos comuns, formando conjuntos do CSC' muito similares.

Tabela 4.16: Alguns termos com alto coeficiente de similaridade obtidos com o CSC' comparando O_1 e O_2

Conceito O_1	Conceito O_2	CSC'
direito constitucional	direito	0.66
direito constitucional	direto eleitoral	0.8
direito eleitoral	direito	0.8

O alto coeficiente retornado pela utilização do CSC' para estes termos, pode servir como base para decisão de um especialista que está realizando o mapeamento semi-automático entre ontologias pois, ao perceber os altos coeficientes de similaridade, o especialista pode concluir que os termos estão em posições similares mas que foram diferentemente representados por uma visão de mundo distinta de quem construiu estas ontologias. A medida SiSe tem com objetivo sinalizar possíveis termos similares, sendo

que a decisão final sobre o mapeamento dos termos fica a cargo de um especialista humano, contribuindo assim para o mapeamento semi-automático entre ontologias.

Após realizado o experimento com a medida MT (ver Seção 4.1.1) e suas adaptações descritas nesta seção, pudemos observar que os melhores resultados obtidos foram com a utilização da adaptação do CSC, denominado CSC', de acordo com a Tabela 4.17. O CSC' resultou em coeficientes de similaridade mais altos de acordo com os termos considerados similares por humanos apresentados na Tabela 4.5, retratando maior veracidade nos resultados. No entanto, esta abordagem também pode encontrar altos coeficientes para termos que não são semanticamente similares (falsos positivos).

Tabela 4.17: Comparativo dos coeficientes de similaridade entre os termos das ontologias O_1 e O_2, utilizando a medida MT (SC e CSC) e SiSe (SC' e CSC')

Conceito O_1	Conceito O_2	MT(SC)	MT(CSC)	$SiSe(SC')$	$SiSe(CSC')$
direito eleitoral	direito eleitoral	0.14	1.0	0.33	1.0
eleição	eleições	0.2	0.33	0.5	1.0
partido político	partidos políticos	0.2	0.33	0.5	1.0
voto	voto	0.5	1.0	0.5	1.0

Empiricamente, estabelecemos que os termos semanticamente similares pela medida MT (SC e CSC) e pelas adaptações da medida SiSe (SC' e CSC'), são aqueles cujo coeficiente de similaridade ficou acima ou igual a 0.7. Desta forma, passamos a usar este limiar e consideramos um par de termos semanticamente similares se o valor do coeficiente for igual ou maior que 0.7.

4.2 Estratégia da medida SiSe

Nesta seção iremos descrever a estratégia utilizada para medir a similaridade semântica entre ontologias em português da medida SiSe proposta neste trabalho. A estratégia descrita também é utilizada para todos os experimentos desta dissertação, inclusive os que utilizam a medida original MT (SC e CSC).

Conforme descrito anteriormente (vide Seção 2.2), o autor Noy em [Noy 2004] propõe cinco características a serem analisadas na construção de ferramentas para o mapeamento automático ou semi-automático entre ontologias. Desta cinco características a medida SiSe inclui duas, que são:

- nomes dos conceitos e descrições em linguagem natural;
- hierarquia das classes (relacionamento de subclasse ou hiponímia e superclasse ou hiperonímia).

A seguir descrevemos quatro etapas envolvidas em nossa estratégia para medir a similaridade, que englobam as características descritas acima, conforme a Figura 4.1.

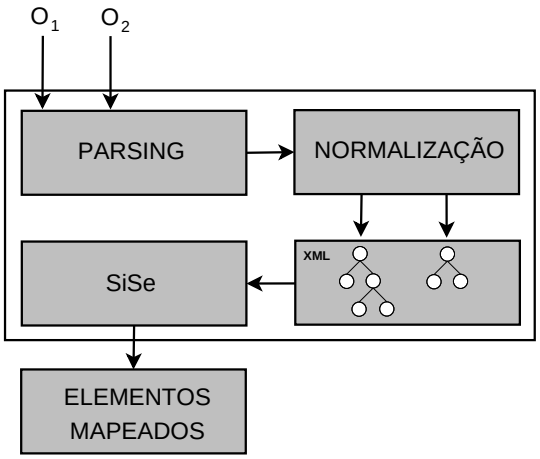


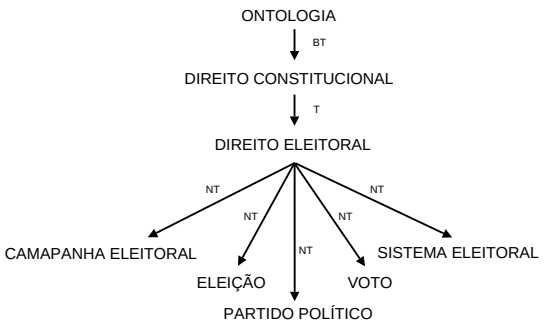
Figura 4.1: Estratégia da medida SiSe

A **primeira etapa** de nossa estratégia consiste na análise da linguagem utilizada na representação das duas ontologias dadas como entrada (O_1 e O_2), como por exemplo, TML (*Thesaurus Markup Language*), RDFS, OWL, DAML. É feito um *parsing* neste conteúdo, extraindo de cada ontologia somente as relações hierárquicas de subconceitos e superconceitos. Por exemplo, a Figura 4.2 (a), representa um trecho de ontologia, descrita com uso da linguagem TML. De acordo com o *parsing* realizado nesta linguagem as relações de hierarquia estão representadas pelas seguintes marcações: T: (linha 03) representa um termo da ontologia; BT: (linha 04), do inglês *Broader Term*, que representa um termo mais geral (superconceito) referente a T; NT: (linhas 05 a 09), do inglês *Narrower Term* representando os termos mais específicos (subconceito) de T.

A Figura 4.2 (b) exemplifica o processo de *parsing* feito na linguagem de representação das ontologias, onde são extraídas as relações de hierarquia da ontologia.

```
01. <?xml version="1.0"?>
02. <THESAURUS tipo="VCBS">
03.     <T term="DIREITO ELEITORAL">
04.       <BT term="DIREITO CONSTITUCIONAL"/>
05.       <NT term ="CAMPANHA ELEITORAL"/>
06.       <NT term="ELEIÇÃO"/>
07.       <NT term="PARTIDO POLÍTICO"/>
08.       <NT term="SISTEMA ELEITORAL"/>
09.       <NT  term="VOTO"/>
10.     </T>
11. </THESAURUS>
```

(a)



(b)

Figura 4.2: Etapa de *parsing* (análise da linguagem utilizada na descrição da ontologia (a), e busca das relações de subconceito e superconceito entre os termos (b))

Segundo o autor Noy em [Noy 2004], as ontologias podem conter diferenças quanto à

linguagem de representação, o que pode significar que a expressividade destas linguagens seja diferente devido à diferença de sintaxe. Para Noy, este problema pode ser contornado através de uma normalização, ou seja, transformar as diferentes sintaxes das linguagens de representação das ontologias em uma mesma linguagem de descrição.

A **segunda etapa** executa uma *normalização* para que as diferenças de sintaxe das linguagens não interfiram no processo de mapeamento, transformando as ontologias para uma mesma linguagem de representação. As duas ontologias têm suas estruturas hierárquicas representadas em um formato XML, de acordo com a Figura 4.2 (b). Temos como exemplo a normalização de linguagem da ontologia representada pela Figura 4.2 (a), conforme mostrado na Figura 4.3.

```
01. <?xml version="1.0"?>
02. <ontologia>
03.     <classe> DIREITO CONSTITUCIONAL
04.       <subclasse> DIREITO ELEITORAL
05.         <subclasse> CAMPANHA ELEITORAL </subclasse>
06.         <subclasse> ELEIÇÃO </subclasse>
07.         <subclasse> PARTIDO POLÍTICO </subclase>
08.         <subclasse> SISTEMA ELEITORAL </subclasse>
09.         <subclasse> VOTO </subclasse>
10.       </subclasse>
11.     </classe>
12. </ontologia>
```

Figura 4.3: Formato XML para normalizar sintaxes das linguagens que descrevem as ontologias

A **terceira etapa** computa a medida SiSe entre os termos das ontologias. Através da normalização da sintaxe das linguagens das ontologias, ambas são representadas em linguagem XML que representa suas estruturas hierárquicas. Através desta representação XML é possível fazer a comparação dos termos das ontologias. Para cada termo das duas ontologias são feitos os seguintes passos:

1. ***tokenização***: é feito o reconhecimento dos *tokens* de um termo. Um termo pode ser monopalavra, possui apenas um *token*, ou pode ser multipalavra possuindo mais de um *token*. Por exemplo, o termo monopalavra direito, possui o *token* direito. Já o termo multipalavra direito do trabalho possui três *tokens*: direito, do e trabalho.

2. **remoção *stopwords***: após a identificação dos *tokens* do termo, ocorre a remoção de *tokens* considerados irrelevantes para a comparação da similaridade, chamados *stopwords*. As *stoplists* são listas que contêm termos comuns ou mais gerais e que não mudam a semântica da palavra. Normalmente esta lista é composta por preposições e artigos [Jurafsky e Martin 2000]. Por exemplo, o termo direito do trabalho, possui o *token* do como *stopword*, sendo assim este *token* é removido resultando os *tokens* direito e trabalho.

3. ***stemming***: aplica-se o algoritmo de *stemming* PortugueseStemmer para cada *token* do termo. Temos, por exemplo, os *tokens* da fase anterior direito e trabalho,

aplicando o *stemmer* teremos os respectivos *stems* de cada *token*, direit trabalh.

4. **representação final dos termos**: é feita a união dos *stems* dos *tokens* que representam um termo, sendo assim, o primeiro *stem* é representado com todos seus caracteres minúsculos, e os n *tokens* seguintes tem sua primeira letra em maiúscula. Por exemplo, o termo direito do trabalho, representado pelos *stems* direit e trabalh, é notado como direitTrabalh, e esta é a representação final do termo, o qual será comparado com outro termo de outra ontologia que foi submetido ao mesmo processo.

5. **SiSe**: aplicação da medida de similaridade semântica SiSe entre os termos das duas ontologias.

Finalmente, a **quarta etapa** indica os termos mapeados, ou seja, os termos que obtiveram coeficiente de similaridade semântica acima de um limiar especificado previamente.

4.3 Protótipo

Esta seção descreve as funcionalidades de um protótipo desenvolvido para auxiliar o acompanhamento e a análise dos resultados do processo de mapeamento entre ontologias, obtidos pelas medidas de similaridade nele implementadas. O protótipo foi desenvolvido na linguagem de programação Python na plataforma Linux, e é apresentado na Figura 4.4.

O protótipo incorpora a estratégia da medida SiSe descrita na Seção 4.2. Desta forma, abstrai as sintaxes das diferentes linguagens utilizadas para descrição das ontologias e extrai as relações hierárquicas das mesmas, apresentando-as na forma de hierarquia de conceitos (indicação do número 1 na Figura 4.4).

As ontologias a serem submetidas ao processo de similaridade podem utilizar distintas medidas de similaridade. O protótipo permite que o usuário selecione uma das medidas desejadas e especifique um limiar mínimo para os resultados (indicação do número 2 na Figura 4.4). As medidas de similaridade presentes no protótipo são:

- Mapeamento Taxonômico utilizando SC;
- Mapeamento Taxonômico utilizando CSC;
- SiSe utilizando SC';
- SiSe utilizando CSC'.

A partir da escolha de uma das medidas de similaridade o botão "Medir Similaridade" compara os termos das ontologias utilizando a abordagem da medida escolhida, e seus resultados são apresentados na aba "Mapeamento" (número 3 da Figura 4.4).

O usuário pode a qualquer momento escolher uma das medidas de similaridade e obter os coeficientes entre os termos das ontologias. As abas presentes em "Informações"

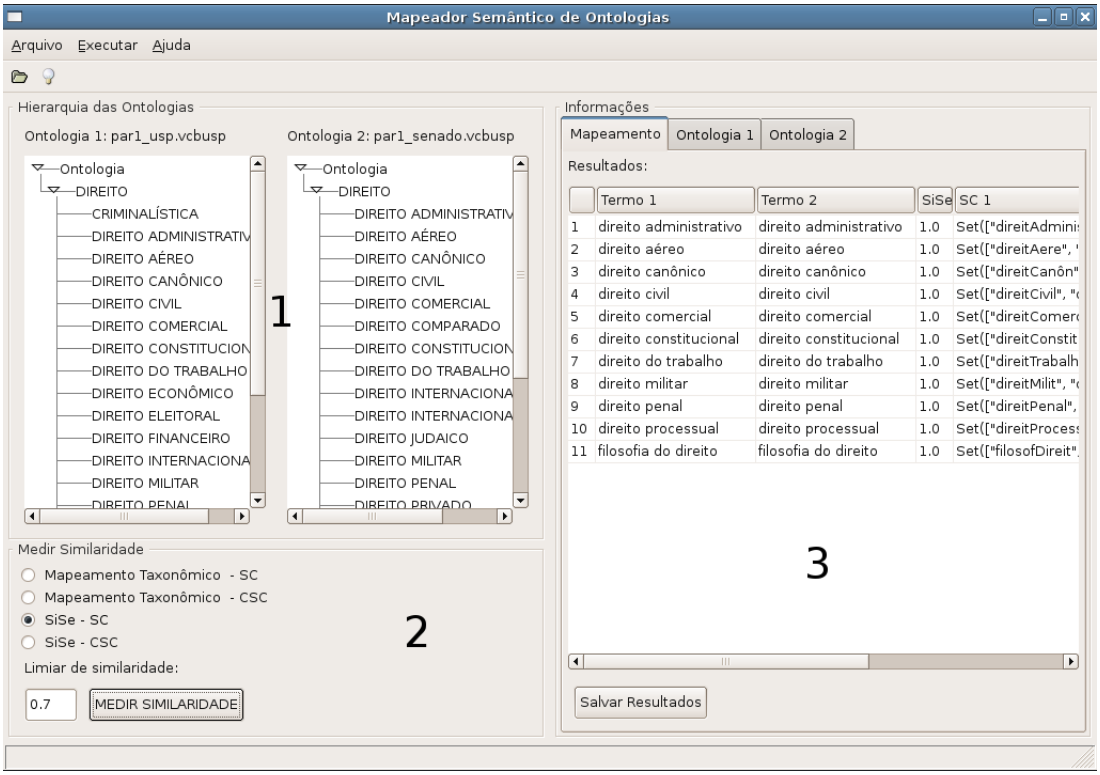


Figura 4.4: Interface do protótipo desenvolvido

têm como objetivo, além de informar os coeficientes de similaridade entre os termos das ontologias, oferecer algumas informações referentes às ontologias como, por exemplo a linguagem original de cada uma delas, bem como a linguagem utilizada pelo protótipo.

4.4 Considerações sobre este capítulo

Este capítulo apresentou primeiramente um experimento com a medida MT descrita na Seção 4.1.1. Foram apresentados exemplos de cálculo desta medida utilizando SC e CSC. Através da comparação de duas hierarquias extraídas de ontologias, apresentamos e comentamos os coeficientes de similaridade resultantes da aplicação desta medida, ressaltando os pontos negativos e positivos e diferenças entre SC e CSC.

A partir destes resultados obtidos com a medida MT, fizemos algumas adaptações (Seção 4.1.2) em relação ao SC e CSC, alterando suas definições através da utilização de um algoritmo de *stemming*, de acordo com a medida SiSe. Fizemos um novo experimento com as mesmas ontologias utilizadas com a medida MT, e apresentamos os resultados da medida SiSe utilizando SC' e CSC' e os comentamos. Pudemos observar que os melhores resultados foram obtidos com a medida SiSe utilizando o CSC'.

Também descrevemos a estratégia de similaridade adotada pela medida SiSe (Seção

4.2), a qual abstrai a sintaxe das linguagens das ontologias extraindo destas as relações de hierarquia. Desta forma utiliza uma representação normalizada XML para comparar os termos das ontologias, aplicando a medida de similaridade SiSe.

Por fim, apresentamos um protótipo que é utilizado para facilitar o processo de mapeamento entre ontologias (Seção 4.3). Desta forma, o usuário que o utiliza tem a facilidade de visualizar a hierarquia das ontologias e escolher entre as medidas de similaridade disponíveis, bem como observar os resultados obtidos. O Capítulo 5 irá apresentar o processo de avaliação da medida SiSe.

Capítulo 5

Avaliação dos Resultados

Após as adaptações realizadas na medida MT, de acordo com a medida SiSe, tivemos como preocupação avaliar os resultados obtidos pela mesma. Através da avaliação pudemos chegar aos pontos positivos e negativos e às adaptações.

Lin [Lin 1998] e Noy [Noy e Musen 2000] relatam que não existe uma maneira padrão para avaliação de medidas de similaridade, e que se trata de uma tarefa muito subjetiva. Os trabalhos correlatos descritos no Capítulo 3 remetem a diferentes formas de avaliação dos resultados de suas abordagens para o mapeamento entre ontologias.

As medidas definidas por Maedche e Staab em [Maedche e Staab 2002], CC e MT, tem como avaliação a análise dos resultados obtidos pela aplicação das mesmas utilizando ontologias do domínio do turismo, construídas por diferentes estudantes em um experimento controlado, comparando estas diferentes versões de ontologias para este domínio. A medida SL, de Chaves [Chaves 2003], adotou uma avaliação que compara os resultados de sua medida com os resultados de uma análise humana. Desta forma, termos de duas ontologias foram julgados similares ou não por humanos, e seus resultados foram comparados com a medida SL. O FCA-Merge [Stumme e Maedche 2001] e o Prompt [Noy e Musen 2003] avaliam seus resultados através da comparação dos resultados com outras ferramentas de mesmo propósito existentes, ferramentas estas utilizadas por humanos, consistindo em um processo semi-automático com a ajuda de um especialista.

A CS [Giunchiglia e Shvaiko 2004], Cupid [Madhavan, Bernstein e Rahm 2001] e Mafra [Maedche *et al.* 2002] tem como estratégia de avaliação de seus resultados a comparação com ferramentas e abordagens ditas de características comuns, analisando os resultados que cada uma obteve.

Analisando estas iniciativas de avaliações, notamos que a comparação com outras abordagens seria difícil devido ao fato destas serem desenvolvidas e avaliadas para ontologias em outras línguas (por exemplo, inglês e alemão). Desta maneira o mais natural para proceder à avaliação foi escolher a comparação de resultados entre os coeficientes das medidas SiSe e MT com uma análise humana. Para tal, empregamos uma metodologia que levou a um "*Golden Mapping*" (GM), a qual é descrita na Seção 5.2. As ontologias utilizadas no processo de avaliação da medida SiSe estão descritas a seguir.

5.1 Ontologias utilizadas

Como a medida SiSe é voltada para ontologias em português nos preocupamos na procura de repositórios de ontologias nesta língua, similar ao repositório de ontologias em inglês do projeto DAML[1].

Percebemos que são escassos os repositórios de ontologias em português descritas nas linguagens padrão para a Web Semântica (RDFS, OWL, DAML, etc), o que dificulta a avaliação de nossa medida. As ontologias utilizadas em nossa avaliação consistem em coleções de vocabulários ou de conceitos, associados de forma explícita ou por meio de relações semânticas, denominadas "estruturas ontológicas". Estas estruturas impossibilitam a criação de medidas de similaridade que analisem propriedades e instâncias, características que poderiam ser analisadas segundo Noy [Noy 2004], para o mapeamento entre ontologias.

As ontologias utilizadas nos experimentos e na avaliação desta dissertação foram as seguintes:

- **Vocabulário Controlado Básico do Senado Federal (VCBS)**: é um Tesauro composto de uma lista de palavras cobrindo diferentes áreas do conhecimento, e é utilizada pelos profissionais da Biblioteca do Senado Federal na catalogação do material existente em sua biblioteca. É especialmente interessante para estabelecer relações mentais entre conceitos que estão presentes nas leis da Constituição Federal Brasileira. Este Tesauro pode ser acessado em http://webthes.senado.gov.br/thes/default-vcbs.htm.

- **Vocabulário Controlado da USP (VCUSP)**: contém uma grande quantidade de conceitos utilizados na indexação de documentos de dados bibliográficos. O vocabulário descrito neste Tesauro abrange várias áreas de conhecimento, utilizando relações de equivalência e hierarquia. Estes conceitos podem ser consultados *on-line* através do site http://143.107.73.99/Vocab/SIBIX652.dll/Index.

Estas duas ontologias são bem gerais e possuem diversos domínios do conhecimento como, por exemplo: medicina, nutrição, agricultura, ciências agrárias, direito, entre outros. Na avaliação da medida SiSe utilizamos trechos do domínio do direito, contidos nestas duas ontologias. A seguir descrevemos a metodologia de avaliação utilizada, chamada *Golden Mapping*.

5.2 *Golden Mapping*

Nosso trabalho tem como estratégia de avaliação a comparação dos resultados obtidos na medida SiSe com uma análise humana cujos resultados denominamos "*Golden Mapping*" (GM), ou "mapeamento dourado". Este tipo de avaliação é amplamente utilizado em PLN. Por exemplo, o recente esforço denominado HAREM, que foi uma avaliação

[1]Este repositório de ontologias possui uma grande variedade de ontologias em diversos domínios. Pode ser acessado em http://www.daml.org/ontologies/

conjunta de sistemas de reconhecimento de entidades mencionadas [Cardoso 2006], fez uso dessa estratégia de avaliação.

A avaliação humana é realizada antes da obtenção dos resultados da medida SiSe. Este fato faz com que a avaliação seja menos tendenciosa, pois os humanos podem encontrar mapeamentos que as medidas não encontram. O GM gera um consenso dos mapeamentos de acordo com as avaliações dos humanos, podendo ser confrontado com os resultados das medidas automáticas para o mapeamento entre ontologias.

Foi estabelecido, então, que a avaliação humana deveria ser feita por três indivíduos distintos. Destes três, definimos as características que cada um deveria ter para realizar a avaliação e os motivos para as escolhas. Estas características se referem à formação.

1. **Lingüista**: este profissional foi escolhido para participar da avaliação por se tratar de um especialista no estudo da linguagem, trazendo importante conhecimento das relações semânticas que estão presentes nas estruturas ontológicas. Desta forma convidamos uma estudante de doutorado em Lingüística para participar da avaliação.

2. **Bacharel em Ciência da Computação**: a escolha deste profissional se deve ao conhecimento adquirido sobre os conceitos de ontologias na Ciência da Computação. Convidamos um mestrando em Ciência da Computação da área de agentes inteligentes e que faz uso de ontologias em seu trabalho.

3. **Bacharel em Direito**: como as ontologias utilizadas para avaliação foram do domínio do Direito, achamos que um profissional desta área seria importante, pois trata da visão de um especialista que tem conhecimento do jargão utilizado neste domínio, podendo assim encontrar mapeamentos que os outros humanos não encontram. Convidamos para esta atividade um Bacharel em Direito especialista em Direito Civil.

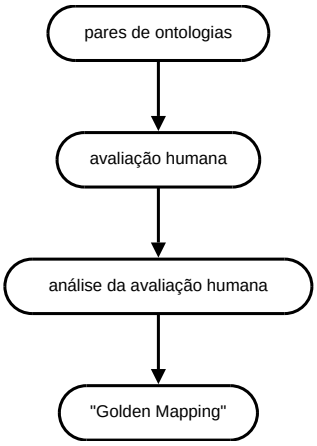


Figura 5.1: Passos para produção do "*Golden Mapping*"

A metodologia de nossa avaliação consiste em quatro passos. Ao final destes passos têm-se uma referência humana para que seja confrontado com o resultado das medidas de similaridade, em nosso caso com a medida SiSe e MT. A seguir temos a descrição destes passos e as orientações que foram dadas aos avaliadores humanos.

- **pares de ontologias**: cada um dos humanos envolvidos no processo de avaliação recebe um documento contendo: (*i*) a estrutura hierárquica dos trechos de ontologias; (*ii*) uma tabela de mapeamento para cada par de ontologias. Nossa avaliação selecionou 5 pares de ontologias. Cada par é formado pelas ontologias VCBS e VCUSP. O documento utilizado pelos humanos para a avaliação dos pares de ontologias está em Anexo H.

- **avaliação humana**: após receber os pares de ontologias, cada avaliador humano indica na tabela de mapeamento os termos que foram considerados similares, assinalando a que ontologia os termos pertencem. Os avaliadores receberam instruções para que a similaridade entre os termos das ontologias seja considerada pela semântica dos mesmos, e não apenas pela representação (combinação de caracteres) do termo. Desta forma, os avaliadores preenchem uma tabela de mapeamento para cada par de extratos de ontologias.

- **análise da avaliação humana**: após realizadas as avaliações pelos humanos nós as analisamos, tentando chegar a um consenso dos termos mapeados pelos três humanos que participaram da avaliação, podendo incluir ou excluir mapeamentos destas pessoas, de acordo com as seguintes regras: será considerado para o GM aquele mapeamento entre os termos das ontologias que foram identificados no mínimo por dois dos três humanos. Sendo que os mapeamentos considerados pelo Bacharel em Direito são sempre considerados, independente da análise dos outros humanos.

- ***Golden Mapping***: ao final deste processo é criada uma referência de mapeamentos entre as ontologias envolvidas. Esta referência de mapeamento é denominada "*Golden Mapping*" e é confrontada com o mapeamento da medida de similaridade.

A Tabela 5.1 apresenta o número de termos que possuem as ontologias de cada par.

5.2.1 Análise do Par 1

A Tabela 5.2 apresenta a hierarquia dos trechos de ontologias do Par 1, do domínio do Direito, onde as mesmas representam o primeiro nível hierárquico deste domínio. Conforme a Tabela 5.1, este par de ontologias possui um total de 44 termos, sendo que 21 termos em O_1 (VCUSP) e 23 termos em O_2 (VCBS).

Este par de ontologias foi submetido a uma avaliação por humanos, como descrito na Seção 5.2. Desta forma cada humano (um total de três) analisou o par de ontologias e identificou os termos que julgou semanticamente similares. A Tabela 5.3 apresenta o número de mapeamentos encontrados pelos humanos, bem como o número de mapeamentos gerados para o GM, onde o Lingüista identificou 16 mapeamentos, o Bacharel em

Tabela 5.1: Informações sobre os pares de ontologias utilizados na avaliação

Ontologias	Número de termos		Total de termos
	VCUSP	VCBS	
Par 1	21	23	44
Par 2	17	16	33
Par 3	20	16	36
Par 4	40	34	74
Par 5	28	19	47
Total	**126**	**108**	**234**

Tabela 5.2: Hierarquias dos trechos de ontologias do Par 1

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 criminalística	02 direito administrativo
03 direito administrativo	03 direito aéreo
04 direito aéreo	04 direito canônico
05 direito canônico	05 direito civil
06 direito civil	06 direito comercial
07 direito comercial	07 direito comparado
08 direito constitucional	08 direito constitucional
09 direito do trabalho	09 direito do trabalho
10 direito econômico	10 direito internacional privado
11 direito eleitoral	11 direito internacional público
12 direito financeiro	12 direito judaico
13 direito internacional	13 direito militar
14 direito militar	14 direito penal
15 direito penal	15 direito privado
16 direito previdenciário	16 direito processual
17 direito processual	17 direito público
18 direito tributário	18 direito romano
19 direito urbanístico	19 filosofia do direito
20 filosofia do direito	20 fontes do direito
21 história do direito	21 jurisprudência
	22 sociologia jurídica
	23 teoria do direito

Ciência da Computação 14 mapeamentos, enquanto o Bacharel em Direito mapeou 13 termos. Os mapeamentos identificados pelos humanos para este par de ontologias estão no Apêndice A.

Tabela 5.3: Número de mapeamentos da análise humana e *Golden Mapping* gerado para o Par 1

Análise humana	Número de mapeamentos
Lingüsta	16
Bacharel em Ciência da Computação	14
Bacharel em Direito	13
Golden Mapping	**14**

Neste par podemos destacar o mapeamento entre os termos direito em O_1 e direito em O_2. Este mapeamento não foi identificado pelo Lingüista e pelo Bacharel em Direito, somente pelo Bacharel em Ciência da Computação.

Acredita-se que o não mapeamento destes termos por estes dois humanos seja pelo fato de tais termos serem as raízes das ontologias. Desta maneira os humanos não julgaram necessário o seu mapeamento. No entanto estes termos foram considerados no GM por serem similares.

Também destacamos os mapeamentos identificados somente pelo Lingüista, que foram: filosofia do direito em O_1 e teoria do direito em O_2, história do direito em O_1 e fontes do direito em O_2 e história do direito em O_1 e filosofia do direito em O_2. Estes termos não fizeram parte do GM, pois somente um dos três humanos identificou a similaridade dos mesmos. O GM gerado para este par de ontologias é constituído de 14 mapeamentos, os quais estão representados pelos termos da Tabela 5.13.

5.2.2 Análise do Par 2

Na Tabela 5.5 apresentamos a hierarquia dos trechos de ontologias do Par 2, onde as mesmas representam o primeiro nível hierárquico do termo direito comercial. Conforme a Tabela 5.1, este par possui um total de 33 termos, 17 destes termos são pertencentes a O_1 (VCUSP) e 16 termos a O_2 (VCBS).

A Tabela 5.4 apresenta o número de mapeamentos encontrados por cada humano envolvido no processo de análise da similaridade entre os termos das duas ontologias, bem como GM gerado a partir desta análise humana, onde o Lingüista encontrou 9 mapeamentos, o Bacharel em Ciência de Computação identificou 8 mapeamentos e o Bacharel em Direito 10 mapeamentos. Os mapeamentos encontrados pelos humanos para este par de ontologias estão no Apêndice B.

Tabela 5.4: Número de mapeamentos da análise humana e *Golden Mapping* gerado para o Par 2

Análise humana	Número de mapeamentos
Lingüista	9
Bacharel em Ciência da Computação	8
Bacharel em Direito	10
Golden Mapping	**11**

Assim como no Par 1, no Par 2 os termos direito em O_1 e direito em O_2 foram mapeados somente pelo Bacharel em Ciência da Computação. No entanto, este mapeamento foi incluído no GM devido à similaridade entre os termos.

O Lingüista identificou três mapeamentos que não foram considerados pelos outros dois humanos. São eles: associação comercial em O_1 e sociedade comercial em O_2, mercadoria em O_1 e mercado de capitais em O_2 e direito falimentar em O_1 e concordata em O_2. Estes mapeamentos, por terem sido identificados somente por um dos humanos, não foram considerados para o GM.

O Bacharel em Direito encontrou três mapeamentos que foram considerados no GM, e que não foram identificados pelos outros humanos. São eles: ato de comércio em O_1 e compra e venda em O_2, contrato comercial em O_1 e compra e venda em O_2 e direito empresarial em O_1 e direito comercial em O_2.

O GM deste par é constituído de 11 mapeamentos, representados pelos termos da Tabela 5.15.

Tabela 5.5: Hierarquias dos trechos de ontologias do Par 2

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito comercial	02 direito comercial
03 associação comercial	03 ato de comércio
04 ato de comércio	04 compra e venda
05 código de proteção e	05 concordata
defesa do consumidor	06 contrato comercial
06 contrato comercial	07 contrato de transporte
07 direito aeronáutico	08 dano
08 direito alfandegário	09 direito autoral
09 direito bancário	10 direito bancário
10 direito cambiário	11 direito cambiário
11 direito da informática	12 direito industrial
12 direito industrial	13 garantia
13 direito empresarial	14 mercado de capitais
14 direito falimentar	15 seguro
15 direito marítimo	16 sociedade comercial
16 mercadoria	
17 sociedade comercial	

5.2.3 Análise do Par 3

O terceiro par de ontologias submetidas a análise humana está representado na Tabela 5.7, onde é apresentada a estrutura hierárquica dos trechos das ontologias. Em ambas as ontologias está representado o primeiro nível hierárquico do termo direito administrativo. Estas ontologias possuem um total de 36 termos sendo que, destes, 20 termos pertencem

a O_1 (VCUSP) e 16 termos a O_2 (VCBS), vide Tabela 5.1.

A Tabela 5.6 apresenta o número de mapeamentos encontrados pelos humanos para o par de ontologias representadas na Tabela 5.7. Para este par, o Lingüista encontrou 15 mapeamentos, enquanto o Bacharel em Ciência da Computação encontrou 9 mapeamentos. Já o Bacharel em Direito encontrou 11 mapeamentos. Os mapeamentos identificados pelos humanos neste par de ontologias estão no Apêndice C.

Tabela 5.6: Número de mapeamentos da análise humana e *Golden Mapping* gerado para o Par 3

Análise humana	Número de mapeamentos
Lingüista	15
Bacharel em Ciência da Computação	9
Bacharel em Direito	11
Golden Mapping	**13**

Tabela 5.7: Hierarquias dos trechos de ontologias do Par 3

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito administrativo	02 direito administrativo
03 administração pública	03 competência administrativa
04 ato administrativo	04 contencioso administrativo
05 competência administrativa	05 direito disciplinar
06 contencioso administrativo	06 direito financeiro
07 contrato administrativo	07 direito tributário
08 controle admisnistrativo	08 domínio público
09 delegação de competência	09 funcionário público
10 direito disciplinar	10 organização administrativa
11 domínio público	11 poder administrativo
12 função administrativa	12 poder de polícia
13 função pública	13 processo administrativo
14 imprevisibilidade	14 reversão
15 jurisdição administrativa	15 serviço público
16 moralidade administrativa	16 ato administrativo
17 processo administrativo	
18 poder administrativo	
19 responsabilidade administrativa	
20 tribunal administrativo	

Podemos destacar o mapeamento de dois pares de termos por parte do Bacharel em Direito. São eles: função pública em O_1 e serviço público em O_2 e, contencioso administrativo em O_1 e processo administrativo em O_2. Estes termos foram considerados no GM.

O Lingüista identificou 4 mapeamentos que não foram considerados para o GM deste par, são eles: função pública em O_1 e funcionário público em O_2, delegação de competência em O_1 e competência administrativa em O_2, jurisdição administrativa em O_1 e poder administrativo em O_2 e, por fim, responsabilidade administrativa em O_1 e competência administrativa em O_1. A análise humana gerou um GM de 13 mapeamentos, conforme os termos apresentados na Tabela 5.17.

5.2.4 Análise do Par 4

A Tabela 5.9 apresenta a estrutura hierárquica dos trechos de ontologias do Par 4. É apresentada a hierarquia do termo direito eleitoral em ambas as ontologias. As ontologias apresentadas possuem um total de 74 termos, constituindo 40 termos em O_1 (VCUSP) e 34 termos em O_2 (VCBS), de acordo com a Tabela 5.1.

A Tabela 5.8 apresenta o número de mapeamentos encontrados pelos humanos para este par de ontologias e também apresenta o número de mapeamentos gerados para o GM. Foram encontrados 22 mapeamentos pelo Lingüista, 22 mapeamentos pelo Bacharel em Ciência da Computação e 18 mapeamentos pelo Bacharel em Direito. Os mapeamentos identificados pelos humanos para este par estão no Apêndice D.

Tabela 5.8: Número de mapeamentos da análise humana e *Golden Mapping* gerado para o Par 4

Análise humana	Número de mapeamentos
Lingüista	22
Bacharel em Ciência da Computação	22
Bacharel em Direito	18
Golden Mapping	**21**

Podemos destacar o mapeamento dos termos votação em O_1 e voto em O_2 identificado pelo Bacharel em Ciência da Computação. Pelo fato destes termos terem sido identificados somente por um humano, os mesmos não foram incluídos no GM. O mesmo ocorreu com os termos justiça eleitoral em O_1 e sistema eleitoral em O_2 que foram identificados pelo Lingüista e, que também não foram considerados no GM. O GM gerado para este par é constituído de 21 mapeamentos, representados na Tabela 5.19.

5.2.5 Análise do Par 5

O quinto par utilizado na avaliação para criação do GM está representado pelas hierarquias apresentadas na Tabela 5.10. As ontologias apresentam a estrutura hierárquica dos termos relacionados a direito internacional. Este par de ontologias é composto de 47 termos, destes, 28 termos em O_1 (VCUSP) e 19 termos em O_2 (VCBS).

Tabela 5.9: Hierarquias dos trechos de ontologias do Par 4

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito eleitoral	02 direito constitucional
03 crime eleitoral	03 direito eleitoral
04 fraude eleitoral	04 campanha eleitoral
05 domicílio eleitoral	05 eleição
06 eleições	06 eleição direta
07 elegibilidade	07 eleição estadual
08 inelegibilidade	08 eleição indireta
09 eleição estadual	09 eleição municipal
10 eleição municipal	10 eleição parlamentar
11 eleição parlamentar	11 eleição presidencial
12 eleição presidencial	12 eleição primária
13 sucessão presidencial	13 partido político
14 eleição primária	14 convenção partidária
15 mandato eletivo	15 partido comunista
16 reeleição	16 partido conservador
17 reforma eleitoral	17 partido democrático
18 justiça eleitoral	18 partido liberal
19 tribunal eleitoral	19 partido republicano
20 tribunal regional federal	20 partido socialista
21 competência eleitoral	21 partido trabalhista
22 partidos políticos	22 sistema eleitoral
23 fidelidade partidária	23 voto
24 fundo partidário	24 voto censitário
25 sistema distrital	25 voto da mulher
26 voto	26 voto distrital
27 cédula eleitoral	27 voto do analfabeto
28 voto censitário	28 voto do menor
29 voto da mulher	29 voto eletrônico
30 voto distrital	30 voto em branco
31 voto do analfabeto	31 voto nulo
32 voto do menor	32 voto obrigatório
33 voto eletrônico	33 voto popular
34 voto em branco	34 voto secreto
35 voto nulo	
36 voto obrigatório	
37 voto popular	
38 voto secreto	
39 votação	
40 contagem de votos	

Tabela 5.10: Hierarquias dos trechos de ontologias do Par 5

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito internacional	02 direito internacional privado
03 direito internacional privado	03 conflito de leis
04 direito comercial internacional	04 sentença estrangeira
05 trips	05 direito internacional público
06 direito econômico internacional	06 arbitragem internacional
07 coisas e bens de direito internacional	07 cláusula de nação mais favorecida
08 concorrência internacional	08 direito de guerra
09 cláusula de nação mais favorecida	09 direito diplomático
10 incoterms	10 direito do mar
11 zona econômica exclusiva	11 direito econômico internacional
12 competência internacional	12 direito fluvial internacional
13 direito iternacional público	13 direito internacional de desenvolvimento
14 direito comunitário	14 direitos humanos
15 direito de guerra	15 jurisdição internacional
16 direito diplomático	16 pessoa jurídica de direito internacional público
17 direito do mar	17 direito consular
18 direito fluvial internacional	18 direito penal internacional
19 direito nuclear	19 relações internacionais
20 direito internacional penal	
21 equilíbrio internacional	
22 estrangeiro	
23 jus gentium	
24 justiça internacional	
25 reconhecimento internacional	
26 represália internacional	
27 sociedade internacional	
28 tratados internacionais	

A Tabela 5.11 apresenta o número de mapeamentos identificados pelos humanos e o GM gerado a partir destes. Foram identificados 15 mapeamentos pelo Lingüista, 14 mapeamentos pelo Bacharel em Ciência da Computação e 11 mapeamentos pelo Bacharel

em Direito. Os mapeamentos identificados pelos humanos para este par de ontologias estão no Apêndice E.

Tabela 5.11: Número de mapeamentos da análise humana e *Golden Mapping* gerado para o Par 5

Análise humana	Número de mapeamentos
Lingüista	15
Bacharel em Ciência da Computação	14
Bacharel em Direito	11
Golden Mapping	**15**

Da análise humana podemos destacar os mapeamentos encontrados pelo Lingüista, que identificou os termos represália internacional em O_1 e direito penal internacional em O_2 como sendo termos similares entre as ontologias. Este mapeamento não foi considerado pelo GM, pois somente um humano o indicou. O mapeamento identificado pelo Bacharel em Ciência da Computação, não incluído no GM, foi dos termos zona econômica exclusiva em O_1 e jurisdição internacional em O_2. Já os termos justiça internacional em O_1 e jurisdição internacional em O_2 foram considerados similares pelo Bacharel em Direito. Desta forma este mapeamento foi incluído no GM. O GM deste par é formado por 15 mapeamentos representados pelos termos da Tabela 5.21.

5.3 Avaliação SiSe x *Golden Mapping*

Após descrita a metodologia de avaliação e a análise feita pelos humanos para os pares de ontologias selecionados, avaliamos os resultados obtidos com a medida de similaridade SiSe e confrontamos com os mapeamentos gerados pelo GM, bem como os resultados obtidos com a medida MT. Através do protótipo descrito na Seção 4.3 aplicamos as medidas SiSe (SC' e CSC') e MT (SC e CSC) e, comentamos a seguir os mapeamentos identificados por estas, de acordo com o limiar 0.7, ou seja, os termos mapeados pelas medidas são aqueles maiores ou iguais ao valor 0.7.

5.3.1 Avaliação do Par 1

O Par 1 de ontologias está representado na Tabela 5.2. Utilizamos o protótipo para mapeamento entre ontologias descrito na Seção 4.3, desta forma obtivemos os coeficientes de similaridade das medidas SiSe (SC' e CSC') e MT (SC e CSC).

De acordo com a Tabela 5.12, dos 14 mapeamentos identificados pelo GM as medidas MT e SiSe, utilizando SC e SC' respectivamente, mapearam 11 termos cada uma, constituindo 78.57% dos mapeamentos. Utilizando CSC e CSC', foram identificados 12 mapeamentos por cada umas das medidas, em um total de 85.71% dos mapeamentos possíveis de acordo com o GM.

Todos os mapeamentos encontrados pelas medidas SiSe e MT obtiveram coeficientes de similaridade igual a 1.0. Este alto coeficiente entre os termos mapeados se deve à

similaridade entre as hierarquias das ontologias.

Os termos das ontologias deste par, ao serem comparados, possuem conjuntos muito pequenos formados pelas medidas SiSe e MT pois, os mesmos possuem poucos subconceitos. O alto coeficiente também se dá pela similaridade lexical entre os termos. A Tabela 5.13 apresenta os coeficientes de similaridade obtidos com as medidas SiSe e MT para os termos mapeados pelo GM deste par. São considerados similares aqueles termos cujo o coeficiente retornou valor igual ou maior que 0.7 (valores em negrito na Tabela 5.13).

Tabela 5.12: Número de mapeamentos identificados pelas medidas MT (SC e CSC) e SiSe (SC' e CSC') para o Par 1, de acordo com o GM

	Valor Absoluto	Valor Percentual
Quantidade total de mapeamentos (GM)	14	100%
Quantidade de mapeamentos MT(SC)	11	78.57%
Quantidade de mapeamentos MT(CSC)	12	85.71%
Quantidade de mapeamentos SiSe(SC')	11	78.57%
Quantidade de mapeamentos SiSe(CSC')	12	85.71%

Tabela 5.13: Coeficientes de similaridade das medidas MT e SiSe entre os termos mapeados pelo GM para o Par 1

	Termo O_1	Termo O_2	SC	CSC	SC'	CSC'
01	direito	direito	0.37	**1.0**	0.37	**1.0**
02	direito administrativo	direito administrativo	**1.0**	**1.0**	**1.0**	**1.0**
03	direito aéreo	direito aéreo	**1.0**	**1.0**	**1.0**	**1.0**
04	direito canônico	direito canônico	**1.0**	**1.0**	**1.0**	**1.0**
05	direito civil	direito civil	**1.0**	**1.0**	**1.0**	**1.0**
06	direito comercial	direito comercial	**1.0**	**1.0**	**1.0**	**1.0**
07	direito constitucional	direito constitucional	**1.0**	**1.0**	**1.0**	**1.0**
08	direito do trabalho	direito do trabalho	**1.0**	**1.0**	**1.0**	**1.0**
09	direito internacional	direito internacional privado	0.33	0.33	0.33	0.33
10	direito internacional	direito internacional público	0.33	0.33	0.33	0.33
11	direito militar	direito militar	**1.0**	**1.0**	**1.0**	**1.0**
12	direito penal	direito penal	**1.0**	**1.0**	**1.0**	**1.0**
13	direito processual	direito processual	**1.0**	**1.0**	**1.0**	**1.0**
14	filosofia do direito	filosofia do direito	**1.0**	**1.0**	**1.0**	**1.0**

Os termos direito em O_1 e direito em O_2 resultaram um coeficiente de similaridade baixo de acordo com as medidas MT e SiSe utilizando SC e SC' respectivamente. Estes termos resultaram em um coeficiente 0.37. O baixo coeficiente entre estes dois termos se deve à diferença de hierarquia entre os mesmos, onde o termo direito em O_1 possui um número menor de subconceitos do que o termo direito em O_2. No entanto, estes termos foram considerados similares pelas medidas SiSe e MT utilizando o CSC' e CSC respectivamente, com um coeficiente 1.0 para ambos. Este mapeamento identificado pelo CSC e CSC' se deve ao fato de os mesmos considerarem somente os termos comuns em ambas as ontologias para os subconceitos destes termos.

Já os termos identificados como similares pelo GM como, por exemplo direito internacional em O_1 e direito internacional privado em O_2 e, os termos direito internacional em

O_1 e direito internacional público em O_2 não foram identificados como similares pelas medidas, em todas o coeficiente ficou em 0.33. Neste caso a baixa similaridade entre estes termos se deve à diferença do número de palavras que compõem cada um, onde direito internacional em O_1 é composto de duas palavras, enquanto direito internacional privado e direito internacional público são constituídos por três palavras. Estes são considerados como distintos nos conjuntos formados pelas medidas SiSe e MT, baixando o coeficiente de similaridade.

5.3.2 Avaliação do Par 2

A Tabela 5.5 apresenta as ontologias do Par 2. Utilizamos este par de ontologias em nosso protótipo e obtivemos os coeficientes de similaridade entre os termos destas ontologias.

Conforme a Tabela 5.14 temos 11 mapeamentos identificados pelo GM. As medidas SiSe e MT utilizando o SC' e SC identificaram cada uma 6 mapeamentos, constituindo 54.54% dos termos mapeados pelo GM. Já utilizando o CSC' e CSC identificaram como similares 8 termos cada um, o que significa 72.72% dos termos do GM.

Tabela 5.14: Número de mapeamentos identificados pelas medidas MT (SC e CSC) e SiSe (SC' e CSC') para o Par 2, de acordo com o GM

	Valor Absoluto	Valor Percentual
Quantidade total de mapeamentos (GM)	11	100%
Quantidade de mapeamentos MT(SC)	6	54.54%
Quantidade de mapeamentos MT(CSC)	8	72.72%
Quantidade de mapeamentos SiSe(SC')	6	54.54%
Quantidade de mapeamentos SiSe(CSC')	8	72.72%

Todos os mapeamentos identificados pelas medidas SiSe e MT para os termos do GM tiveram como coeficiente 1.0. Este alto coeficiente se explica pelo fato das hierarquias das ontologias serem muito similares, bem como estes termos comparados serem lexicalmente idênticos. A Tabela 5.15 apresenta os coeficientes de similaridade dos termos do GM.

Tabela 5.15: Coeficientes de similaridade das medidas MT e SiSe entre os termos mapeados pelo GM para o Par 2

	Termo O_1	Termo O_2	SC	CSC	SC'	CSC'
01	direito	direito	0.32	**1.0**	0.32	**1.0**
02	direito comercial	direito comercial	0.32	**1.0**	0.32	**1.0**
03	ato de comércio	ato de comércio	**1.0**	**1.0**	**1.0**	**1.0**
04	contrato comercial	contrato comercial	**1.0**	**1.0**	**1.0**	**1.0**
05	direito bancário	direito bancário	**1.0**	**1.0**	**1.0**	**1.0**
06	direito cambiário	direito cambiário	**1.0**	**1.0**	**1.0**	**1.0**
07	direito industrial	direito industrial	**1.0**	**1.0**	**1.0**	**1.0**
08	sociedade comercial	sociedade comercial	**1.0**	**1.0**	**1.0**	**1.0**
09	ato de comércio	compra e venda	0.5	0.5	0.5	0.5
10	contrato comercial	compra e venda	0.5	0.5	0.5	0.5
11	direito empresarial	direito comercial	0.12	0.22	0.12	0.22

Os termos direito em O_1 e direito em O_2 e direito comercial em O_1 e direito comercial em O_2 não foram considerados similares pelas medidas SiSe e MT utilizando SC' e SC respectivamente, obtiveram um coeficiente 0.32. Isto se deve à diferença no número de subconceitos entre os termos, onde os termos em O_1 possuem um maior número de subconceitos do que os termos em O_2, diminuindo a similaridade de acordo com suas hierarquias. Estes mesmos termos foram identificados como similares pelo CSC' e CSC, com coeficiente 1.0, pelo fato dos mesmos considerarem somente os termos similares em ambas as ontologias como elementos de seus conjuntos.

Já os termos ato de comércio em O_1 e compra e venda em O_2 e, os termos contrato comercial em O_1 e compra e venda em O_2 não foram mapeados por nenhuma das medidas do protótipo e obtiveram coeficiente 0.5 de todas as medidas. Isto se deve à diferença lexical dos termos em questão, visto que seus superconceitos são os mesmos (direito e direito comercial para ambos os pares de termos). Os termos direito empresarial em O_1 e direito comercial em O_2 têm um coeficiente baixo em todas as medidas por estarem em posições hierárquicas distintas, donde o termo direito comercial em O_2 possui um maior número de subconceitos que o termo direito empresarial em O_1, baixando o coeficiente de similaridade.

Além dos 8 mapeamentos identificados pelo CSC' e CSC pertencentes ao GM, foram mapeados ainda por ambos, mais dois termos, são eles: direito em O_1 e direito comercial em O_2, e direito comercial em O_1 e direito em O_2, com um coeficiente igual a 1.0 para ambos. Estes termos foram dados como similares pelas medidas, no entanto não pertencem ao termos mapeados pelo GM para este par de ontologias, sendo assim são chamados de **falsos positivos**.

5.3.3 Avaliação do Par 3

O Par 3 avaliado nesta dissertação está representado pelas hierarquias das ontologias da Tabela 5.7. Utilizamos o protótipo para medir a similaridade entre os termos destas ontologias.

A Tabela 5.16 apresenta o número de mapeamentos identificados pelas medidas de similaridade de acordo com o GM. Desta forma, é visto que o MT utilizando SC e a SiSe utilizando SC' mapearam 7 termos cada um, significando 53.84% dos mapeamentos do GM. Já o CSC e CSC' mapearam 9 termos cada um, o que representa 69.23% dos mapeamentos do GM.

Tabela 5.16: Número de mapeamentos identificados pelas medidas MT (SC e CSC) e SiSe (SC' e CSC') para o Par 3, de acordo com o GM

	Valor Absoluto	Valor Percentual
Quantidade total de mapeamentos (GM)	13	100%
Quantidade de mapeamentos MT(SC)	7	53.84%
Quantidade de mapeamentos MT(CSC)	9	69.23%
Quantidade de mapeamentos SiSe(SC')	7	53.84%
Quantidade de mapeamentos SiSe(CSC')	9	69.23%

Todos os termos identificados como similares pelas medidas de similaridade e que fazem parte do GM tiveram coeficiente de similaridade igual 1.0. Assim como os pares de ontologias anteriores, este par também possui hierarquias muito similares, bem como termos lexicalmente similares. A Tabela 5.17 apresenta os coeficientes gerados pelas medidas para os termos do GM.

Tabela 5.17: Coeficientes de similaridade das medidas MT e SiSe entre os termos mapeados pelo GM para o Par 3

	Termo O_1	Termo O_2	SC	CSC	SC'	CSC'
01	direito	direito	0.33	**1.0**	0.33	**1.0**
02	direito administrativo	direito administrativo	0.33	**1.0**	0.33	**1.0**
03	ato administrativo	ato administrativo	**1.0**	**1.0**	**1.0**	**1.0**
04	competência administrativa	competência administrativa	**1.0**	**1.0**	**1.0**	**1.0**
05	contencioso administrativo	contencioso administrativo	**1.0**	**1.0**	**1.0**	**1.0**
06	direito disciplinar	direito disciplinar	**1.0**	**1.0**	**1.0**	**1.0**
07	domínio público	domínio público	**1.0**	**1.0**	**1.0**	**1.0**
08	poder administrativo	poder administrativo	**1.0**	**1.0**	**1.0**	**1.0**
09	controle administrativo	poder administrativo	0.5	0.5	0.5	0.5
10	processo administrativo	processo administrativo	**1.0**	**1.0**	**1.0**	**1.0**
11	função pública	serviço público	0.5	0.5	0.5	0.5
12	jurisdição administrativa	competência administrativa	0.5	0.5	0.5	0.5
13	contencioso administrativo	processo administrativo	0.5	0.5	0.5	0.5

Os termos direito em O_1 e direito em O_2 e direito administrativo em O_1 e direito administrativo em O_2 não foram considerados similares pelo SC e SC', devido a diferença na hierarquia dos termos, onde em O_1 possuem um maior número de subconceitos do que em O_2. No entanto, as medidas utilizando CSC e CSC' deram estes termos como similares pois consideram somente os superconceitos e subconceitos similares em ambas as ontologias, o que resultou em um aumento no coeficiente em relação ao SC e SC'.

Alguns termos por terem representação lexical distintas não ficaram com coeficiente de similaridade 0.7. Dentre eles, destacamos os termos pertencentes ao GM: controle administrativo em O_1 e poder administrativo em O_2; função pública em O_1 e serviço público em O_2; jurisdição administrativa em O_1 e competência administrativa em O_2; e contencioso administrativo em O_1 e processo administrativo em O_2. Apesar destes termos terem os mesmos superconceitos (direito e direito administrativo) são distintos lexicalmente o que diminui o coeficiente de similaridade entre os mesmos.

Além dos 9 mapeamentos identificados através do CSC' e CSC pertencentes ao GM, também foram mapeados (falsos positivos) os termos direito em O_1 e direito administrativo em O_2 e direito administrativo em O_1 e direito em O_2, com um coeficiente igual a 1.0.

5.3.4 Avaliação do Par 4

As ontologias do Par 4 estão representadas pelas hierarquias da ontologias da Tabela 5.9. Foi utilizado o protótipo descrito na Seção 4.3 para obtenção dos coeficientes de similaridade entre os termos destas ontologias.

A Tabela 5.18 apresenta os mapeamentos obtidos pelas medidas SiSe e MT de acordo com o GM. Desta forma, o MT utilizando o SC identificou 12 mapeamentos, constituindo

57.14% dos mapeamentos possíveis. Já o CSC identificou 20 mapeamentos, o equivalente a 95.23% dos mapeamentos do GM. A medida SiSe utilizando SC' identificou 16 mapeamentos, constituindo 76.19% dos mapeamentos. Utilizando o CSC' foram encontrados 100% (21) dos mapeamentos de acordo com o GM.

Tabela 5.18: Número de mapeamentos identificados pelas medidas MT (SC e CSC) e SiSe (SC' e CSC') para o Par 4, de acordo com o GM

	Valor Absoluto	Valor Percentual
Quantidade total de mapeamentos (GM)	21	100%
Quantidade de mapeamentos MT(SC)	12	57.14%
Quantidade de mapeamentos MT(CSC)	20	95.23%
Quantidade de mapeamentos SiSe(SC')	16	76.19%
Quantidade de mapeamentos SiSe(CSC')	21	100%

A Tabela 5.19 apresenta os coeficientes de similaridade para os termos do GM. Podemos observar um ganho significativo nos coeficiente de similaridade entre os termos que utilizaram SC em relação ao SC'. O fato do SC' considerar os *stems* dos termos, faz com que as diferenças lexicais (como, por exemplo, os termos eleições em O_1 e eleição em O_2 e partidos políticos em O_1 e partido político em O_2) tenham um única representação, aumentando o coeficiente de similaridade.

Tabela 5.19: Coeficientes de similaridade das medidas MT e SiSe entre os termos mapeados pelo GM para o Par 4

	Termo O_1	Termo O_2	SC	CSC	SC'	CSC'
01	direito	direito	0.34	**1.0**	0.4	**1.0**
02	direito eleitoral	direito eleitoral	0.34	**1.0**	0.4	**1.0**
03	eleições	eleição	0.38	**0.77**	0.47	**1.0**
04	eleição estadual	eleição estadual	0.5	**1.0**	**0.8**	**1.0**
05	eleição municipal	eleição municipal	0.5	**1.0**	**0.8**	**1.0**
06	eleição parlamentar	eleição parlamentar	0.5	**1.0**	**0.8**	**1.0**
07	eleição presidencial	eleição presidencial	0.42	**1.0**	0.66	**1.0**
08	eleição primária	eleição primária	0.5	**1.0**	**0.8**	**1.0**
09	partidos políticos	partido político	0.13	0.5	0.21	**1.0**
10	voto	voto	**0.77**	**1.0**	**0.82**	**1.0**
11	voto censitário	voto censitário	**0.8**	**1.0**	**0.8**	**1.0**
12	voto da mulher	voto da mulher	**0.8**	**1.0**	**0.8**	**1.0**
13	voto distrital	voto distrital	**0.8**	**1.0**	**0.8**	**1.0**
14	voto do analfabeto	voto do analfabeto	**0.8**	**1.0**	**0.8**	**1.0**
15	voto do menor	voto do menor	**0.8**	**1.0**	**0.8**	**1.0**
16	voto eletrônico	voto eletrônico	**0.8**	**1.0**	**0.8**	**1.0**
17	voto em branco	voto em branco	**0.8**	**1.0**	**0.8**	**1.0**
18	voto nulo	voto nulo	**0.8**	**1.0**	**0.8**	**1.0**
19	voto obrigatório	voto obrigatório	**0.8**	**1.0**	**0.8**	**1.0**
20	voto popular	voto popular	**0.8**	**1.0**	**0.8**	**1.0**
21	voto secreto	voto secreto	**0.8**	**1.0**	**0.8**	**1.0**

Podemos observar que o CSC' encontrou todos os mapeamentos de acordo com o GM, e todos com coeficiente 1.0. Os termos partidos políticos em O_1 e partido político e O_2 foram mapeados somente pelo CSC' devido a representação dos termos por seus

stems. Também podemos observar que o MT utilizando o CSC identificou a maioria dos mapeamentos do GM para este par, apesar das diferenças lexicais entre alguns dos termos deste par.

É importante ressaltar o maior número de mapeamentos identificados pelo SC' em relação ao SC, um aumento de 19.05% dos mapeamentos em relação ao GM. Este aumento se deve a representação dos termos nos conjuntos gerados pelo SC' através de seus *stems*. Também ocorreu um aumento nos coeficientes na maioria dos termos, mesmo aqueles que não foram mapeados (coeficientes de similaridade abaixo de 0.7)

Apesar de ter mapeado quase a totalidade dos termos do GM, o CSC e CSC' também mapearam alguns termos que não fazem parte do GM, os chamados falsos positivos. O CSC identificou um total de 30 mapeamentos, sendo 20 mapeamentos pertencentes ao GM e, 10 mapeamentos considerados falsos positivos encontrados pela medida. Os mapeamentos falsos positivos encontrados pelo CSC estão no Apêndice F. A medida SiSe utilizando o CSC' também encontrou falsos positivos, do total de 47 mapeamentos identificados pela medida, 21 fazem parte do GM e 26 são falsos positivos. Estes mapeamentos considerados falsos positivos são apresentados no Apêndice G. Um fator negativo do CSC e de sua adaptação CSC', para este par foi o excessivo número de falsos positivos encontrados, pelo fato de considerarem somente os termos comuns às ontologias na análise das hierarquias.

5.3.5 Avaliação do Par 5

O Par 5 está representado pelas hierarquias das ontologias da Tabela 5.10. Os termos destas ontologias tiveram sua similaridade comparada através das medidas MT e SiSe utilizando o protótipo descrito nesta dissertação.

De acordo com a Tabela 5.20 podemos observar que o SC e SC' identificaram 4 mapeamentos cada, o que representa 26.66% dos mapeamentos possíveis de acordo com o GM. Já com a utilização do CSC e CSC' foram mapeados 7 termos cada, representando 46.66% do total de possíveis mapeamentos do GM. O baixo número de mapeamentos para este par de ontologias se deve às diferenças nas hierarquias das mesmas.

Tabela 5.20: Número de mapeamentos identificados pelas medidas MT (SC e CSC) e SiSe (SC' e CSC') para o Par 5, de acordo com o GM

	Valor Absoluto	Valor Percentual
Quantidade total de mapeamentos (GM)	15	100%
Quantidade de mapeamentos MT(SC)	4	26.66%
Quantidade de mapeamentos MT(CSC)	7	46.66%
Quantidade de mapeamentos SiSe(SC')	4	26.66%
Quantidade de mapeamentos SiSe(CSC')	7	46.66%

A Tabela 5.21 apresenta os coeficientes de similaridade dos termos mapeados pelo GM. Neste par de ontologias notamos que foram encontrados poucos mapeamentos utilizando SC e SC'. Isto se deve ao fato de os termos comparados terem hierarquias distintas,

diminuindo o coeficiente de similaridade. Outro fator que impediu alguns mapeamentos, foi o número de palavras que compõem os termos como, por exemplo, os termos direito internacional em O_1 e direito internacional privado em O_2, que possuem número de palavras distintos, desta forma mesmo com hierarquias similares diminui o coeficiente de similaridade.

Além dos mapeamentos dos termos do GM, o CSC e CSC' identificaram como similares (falsos positivos) os termos direito em O_1 e direito internacional público em O_2, em ambos com coeficiente de 1.0. Estes termos foram mapeados por terem hierarquias idênticas formadas por termos comuns em ambas as ontologias.

Tabela 5.21: Coeficientes de similaridade das medidas MT e SiSe entre os termos mapeados pelo GM para o Par 5

	Termo O_1	Termo O_2	SC	CSC	SC'	CSC'
01	direito	direito	0.23	**1.0**	0.23	**1.0**
02	direito internacional	direito internacional privado	0.14	0.4	0.14	0.4
03	direito internacional	direito internacional público	0.12	0.3	0.12	0.3
04	direito internacional privado	direito internacional privado	0.28	**1.0**	0.28	**1.0**
05	direito internacional público	direito internacional público	0.22	**1.0**	0.22	**1.0**
06	direito econômico internacional	direito econômico internacional	0.22	0.5	0.22	0.5
07	cláusula de nação mais favorecida	cláusula de nação mais favorecida	0.4	0.5	0.4	0.5
08	direito de guerra	direito de guerra	**1.0**	**1.0**	**1.0**	**1.0**
09	direito diplomático	direito diplomático	**1.0**	**1.0**	**1.0**	**1.0**
10	direito do mar	direito do mar	**1.0**	**1.0**	**1.0**	**1.0**
11	direito fluvial internacional	direito fluvial internacional	**1.0**	**1.0**	**1.0**	**1.0**
12	direito internacional penal	direito penal internacional	0.5	0.5	0.5	0.5
13	justiça internacional	direito penal internacional	0.5	0.5	0.5	0.5
14	tratados internacionais	relações internacionais	0.5	0.5	0.5	0.5
15	justiça internacional	jurisdição internacional	0.5	0.5	0.5	0.5

5.4 Considerações sobre este capítulo

Este capítulo descreveu a avaliação dos resultados obtidos com as medidas de similaridade para o mapeamento entre ontologias. Foram avaliados os resultados das medidas de similaridade MT e de sua adapatação proposta nesta dissertação, a medida SiSe.

Primeiramente escolhemos duas ontologias em português para que trechos das mesmas fossem extraídos e pudessem ser comparados. Desta forma, a Seção 5.1 descreve as estruturas ontológicas utilizadas em nossa avaliação, que são: VCBS e VCUSP.

Após uma breve descrição dos métodos de avalição realizados em trabalhos correlatos, decidimos por fazer uma avaliação baseada na análise humana.

A metodologia de avaliação dos resultados das medidas de similaridade foi a de "*Golden Mapping*", ou "mapeamento dourado". A metodologia foi descrita na Seção 5.2.

Esta metodologia consiste na avaliação humana das ontologias, ou seja, humanos analisam pares de ontologias e indicam os termos entre as ontologias que julgaram similares. Elencamos para participar deste processo de construção de *Golden Mapping* três humanos, de três especialidades distintas, foram eles: um Lingüista, um Bacharel em Ciência da Computação e um Bacharel em Direito. Da análise humana temos o consenso

dos mapeamentos para os pares de ontologias selecionados para esta avaliação. Desta forma comparamos os mapeamentos calculados pelas medidas de similaridade com os mapeamentos identificados pelo GM.

Através dos resultados obtidos com as medidas SiSe e MT vimos que a medida que em geral encontrou o maior número de mapeamentos de acordo com o GM para cada par de ontologias foi a medida SiSe utilizando o CSC'.

Para os pares de ontologias em que as hierarquias são similares, o número de mapeamentos não variou muito entre as medidas. Vimos que, quando as ontologias possuem termos com pequenas diferenças lexicais, as abordagens que utilizam o *stemming* (SC' e CSC') obtêm os melhores resultados. Também percebemos que o CSC e CSC' encontram um maior número de mapeamentos quando as hierarquias das ontologias possuem níveis diferentes, no entanto também encontram um número elevado de falsos positivos.

Capítulo 6

Conclusão

6.1 Sobre o trabalho

Este trabalho teve como objetivo adaptar uma medida de similaridade semântica para o mapeamento entre ontologias em português. Iniciamos com a realização de uma pesquisa bibliográfica que possibilitou um embasamento teórico sobre as medidas de similaridade utilizadas para o mapeamento entre ontologias. A partir da pesquisa realizada, foi possível adaptar uma medida de similaridade da literatura e avaliar os resultados obtidos.

As medidas de similaridade são de vital importância para integrar e gerar novas fontes de conhecimento através de modelos de dados já existentes. Quanto à similaridade entre ontologias, as medidas de similaridade para termos, ou para ontologias com um todo, são o principal processo envolvido no mapeamento, tanto para a união como para o alinhamento. A união de ontologias tem como principal característica a junção de duas ou mais ontologias, criando uma nova ontologia com todas as características expressas nas ontologias envolvidas no processo. Já o alinhamento de ontologias realiza uma correspondência entre os elementos similares das ontologias envolvidas.

A identificação da similaridade entre os termos das ontologias é obtida através de duas abordagens identificadas na literatura. São elas: similaridade lexical e similaridade semântica.

A abordagem para medir a similaridade lexical considera a forma como os elementos das ontologias são representados, ou seja, sua representação através de cadeias de caracteres. A maioria das abordagens usa um coeficiente de similaridade que avalia, lexicalmente ou estruturalmente, estes elementos. Atualmente as abordagens de similaridade lexical são mais indicadas para as aplicações que têm o tempo de resposta como principal preocupação, pois são relativamente rápidas, executam apenas combinações de caracteres entre os termos. Algumas usam técnicas de PLN como algoritmos para redução da palavra (*stemming*), ou usam recursos lingüísticos como léxicos (por exemplo, WordNet) e Tesauros para análise morfológica, entre outros. Todos estes recursos consideram a comparação de cadeias de caracteres e não o que essas cadeias de caracteres significam, dentro de um contexto específico (significado). Isso torna a similaridade lexical pouco indicada para aplicações que possuem termos reunidos em estruturas diferentes, causando

inconsistências semânticas ao integrar ou alinhar uma estrutura ontológica.

A abordagem semântica considera o significado destes termos, bem como as relações semânticas existentes entre eles e, ainda, a estrutura da taxonomia. Também utilizam recursos e técnicas de PLN como, por exemplo, Tesauro, corpora anotados, bases de dados lexicais (WordNet, por exemplo), algoritmos de *stemming* e desambiguação do sentido das palavras, entre outros. A similaridade semântica oferece uma contribuição diferente da similaridade lexical. Considera o significado dos termos envolvidos no processo, e não apenas as cadeias de caracteres que os constituem. As abordagens semânticas, em sua maioria, procuram em um léxico da linguagem por sinônimos dos termos, buscando relações semânticas entre termos, ou analisando a posição do termo em relação à hierarquia como um todo (similaridade semântico-estrutural). Estas abordagens retornam a similaridade entre termos com uma base mais sólida que a abordagem lexical, pois podem considerar a sinonímia ou outras relações semânticas entre os termos. São, portanto, mais coerentes do que as abordagens lexicais, no entanto podem ser mais lentas em sua resposta, e podem não servir para algumas aplicações.

O mapeamento de ontologias é uma área que está sendo muito pesquisada, e muitas ferramentas e métodos estão sendo desenvolvidos nesta área. No entanto, a automação completa deste processo ainda está longe de acontecer, pois ontologias especificam suas conceitualizações fazendo referência a termos descritos em linguagem natural. As abordagens (lexical e semântica) que medem a similaridade entre os termos das ontologias trazem consigo os mesmos problemas enfrentados na área de PLN. Cada contexto (domínio) tem seus termos, e cada engenheiro de ontologias tem sua visão e a expressa de maneira particular e diferente dos outros.

Ambas as abordagens, lexical e semântica, resolvem apenas parcialmente o problema da detecção da similaridade entre termos de ontologias ou entre ontologias com um todo. Muitas vezes é preciso um humano para fazer a detecção da similaridade manualmente. O problema da similaridade entre termos tem sua origem no uso da linguagem natural, que é ambígua e dependente de contexto, permanecendo o mapeamento entre ontologias um processo ainda semi-automático.

Uma lacuna da área, que pretendemos preencher com esta dissertação, é a investigação no que se refere ao mapeamento de ontologias na língua portuguesa. Poucas abordagens testadas ou propostas para nossa língua foram encontradas. Principalmente percebeu-se a falta de abordagens semânticas para ontologias na língua portuguesa, talvez pelo fato de a língua portuguesa não contar ainda com ferramentas e recursos lingüísticos fundamentais para o PLN, como uma base de dados lexical consistente que possa identificar as relações semânticas e significados das palavras da língua. Mas isto não impede que alternativas sejam pesquisadas e criadas, fazendo uso das ferramentas existentes para a língua portuguesa ou para outras línguas.

Tendo em vista o contexto descrito, este trabalho apresentou uma medida de similaridade semântico-estrutural para ontologias em português, denominada **Similaridade Semântica** ou SiSe, que adapta o Mapeamento Taxonômico de Maedche e Staab.

Através da construção de um protótipo foi possível a extração dos coeficientes de similaridade entre os termos das ontologias. Pudemos assim observar e colher os resultados

das medidas SiSe e MT.

Para avaliar os resultados obtidos com essas medidas, utilizamos a metodologia de "*Golden Mapping*". Esta metodologia envolve um consenso da análise da similaridade de ontologias por humanos e a comparação dos coeficientes das medidas com esse mapeamento. Através desta avaliação vimos que as adaptações da medida SiSe (SC' e CSC') obtiveram melhores resultados quando as ontologias possuem diferenças lexicais, devido a utilização do *stemming*.

Os sistemas para a Web Semântica são os principais beneficiários dos processos de mapeamento entre ontologias, uma vez que agentes de software terão que comunicar-se entre si através de suas ontologias, que são distintas e precisam encontrar correspondências entre seus vocabulários. Além disto, estas medidas de similaridade podem ser aplicadas em outras áreas como, por exemplo, banco de dados ou integração de esquemas, bem como auxiliar sistemas de recuperação de informação, verificando a similaridade entre o termo consultado e os termos contidos nos documentos.

6.2 Contribuições

A nosso ver, as principais contribuições deste trabalho foram:

- estudo e revisão bibliográfica sobre o estado da arte das áreas de ontologias, mapeamento entre ontologias e medidas de similaridade envolvidas neste processo;
- adaptação de uma medida de similaridade semântica para ontologias em português, com experimentação e avaliação dos resultados, através da medida SiSe;
- construção de uma estratégia para medir a similaridade entre ontologias, com este ferramental;
- avaliação com o uso de um *Golden Mapping*.

6.3 Limitações

Dentre as limitações desta dissertação pode-se incluir a carência de dados para este, como ontologias em português nas linguagens de marcação semântica (por exemplo, RDFS e OWL). Isto faz com que o desenvolvimento de medidas de similaridade para esta língua fique limitado à análise dos rótulos (cadeias de caracteres) dos termos (classes) e das relações de hierarquia, não permitindo, por exemplo, a análise das definições de propriedades (domínio, abrangência e restrições) e instâncias, entre outros.

Também poderiam ser mencionadas neste trabalho, as limitações referentes a avaliação dos resultados das medidas utilizando o *"Golden Mapping"*. Esta avaliação exigiu um tempo maior do que o previsto tanto no seu preparo, quanto para realização das análises por parte dos humanos. Dois dos humanos que participaram da análise das ontologias para a geração do GM necessitaram de prazo maior que o previsto para análise, o que sugere a idéia de que este tipo de avaliação deve ser executada reunindo todos os

participantes em um experimento controlado. Para a construção do GM obtivemos em alguns casos diferentes respostas referentes a identificação dos mapeamentos por parte dos humanos, fazendo com que regras (vide Seçao 5.2) fossem definidas para a construção do GM.

Quanto à medida SiSe, apesar de ser considerada uma medida semântica por utilizar as relações de hierarquia (subconceitos e superconceitos), a mesma não identifica termos com mesmo sentido mas com representações lexicais distintas. Isto somente é possível se as hierarquias destes termos forem similares (desta forma os mesmos serão considerados similares).

6.4 Trabalhos futuros

Diante dos estudos realizados e dos resultados obtidos podemos deixar como sugestão de continuidade os seguintes trabalhos futuros:

- comparação da medida SiSe com outras medidas de similaridade para o português que forem surgindo na literatura;
- aplicação da medida SiSe a outros idiomas, utilizando o algoritmo de *stemming* específico para a língua utilizada;
- utilização de um léxico ou dicionário de sinônimos para encontrar termos com mesmo significado mas com representações lexicais distintas, aumentando o número de mapeamentos entre as ontologias;
- criação de heurísticas, permitindo a identificação de um maior número de mapeamentos.

Referências

[Antoniou, Franconi e Harmelen 2005]ANTONIOU, G.; FRANCONI, E.; HARMELEN, F. van. Introduction to Semantic Web ontology languages. In: *ReasoningWeb, Proceedings of the Summer School, Malta*, Springer-Verlag, n. 3564, 2005.

[Bechara 2001]BECHARA, E. *Moderna Gramática Portuguesa*. Rio de Janeiro: Editora Lucerna, 2001. 672 p.

[Berners-Lee, Hendler e Lassila 2001]BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. *Scientific American*, May 2001.

[Cardoso 2006]CARDOSO, N. *Avaliação de Sistemas de Reconhecimento de Entidades Mencionadas*. Dissertação (Mestrado) — Faculdade de Engenharia da Universidade do Porto, 2006. 146 p.

[Chandrasekaran, Josephson e Benjamins 1999]CHANDRASEKARAN, B.; JOSEPHSON, R.; BENJAMINS, V. R. What are ontologies, and why do we need them? *IEEE Intelligent Systems*, v. 14, n. 1, p. 20–26, January/February 1999.

[Chaves 2003]CHAVES, M. S. *Mapeamento e comparação de similaridade entre estruturas ontológicas*. Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2003. 93 p.

[Chaves 2003]CHAVES, M. S. Um estudo e apreciação sobre algoritmos de stemming. *In: IX Jornadas Iberoamericanas de Informática, Cartagena de Indias, Colômbia*, 2003.

[Cimiano, Hotho e Staab 2004]CIMIANO, P.; HOTHO, A.; STAAB, S. Clustering concept hierarchies from text. In: *Proceedings of the Conference on Lexical Resources and Evaluation - LREC'2004*, p. 1721–1724, 2004.

[Cimiano, Hotho e Staab 2005]CIMIANO, P.; HOTHO, A.; STAAB, S. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research - JAIR*, v. 24, p. 263–303, August 2005.

[Davies, Fensel e Harmelen 2003]DAVIES, J.; FENSEL, D.; HARMELEN, F. van. *Towards the Semantic Web: ontology-driven knowledge management*. 1^{st} ed. West Sussex: John Wiley & Sons, 2003. 328 p.

[Dellschaft e Staab 2006]DELLSCHAFT, K.; STAAB, S. On how to perform a gold standard based evaluation of ontology learning. In: *Proceedings of the* 5^{th} *International Semantic Web Conference (ISWC)*, 2006.

[Ding e Foo 2002]DING, Y.; FOO, S. Ontology research and development. Part 2 - a review of ontology mapping and evolving. *Journal of Information Science*, v. 28, n. 5, p. 375–388, October 2002.

[Doan e Halevy 2005]DOAN, A.; HALEVY, A. Y. Semantic integration research in the database community: a brief survey. *AI Magazine, Special Issue on Semantic Integration*, v. 26, n. 1, p. 83–94, Spring 2005.

[Edgington *et al.* 2004]EDGINGTON, T. *et al.* Adopting ontology to facilitate knowledge sharing. *Communications of the ACM*, v. 47, n. 11, November 2004.

[Euzenat *et al.* 2004]EUZENAT, J. *et al.* State of the art on ontology alignment. *Knowledge Web Deliverable D2.2.3*, 2004.

[Everett *et al.* 2002]EVERETT, J. O. *et al.* Making ontologies work for resolving redundancies across documents. *Communications of the ACM*, v. 45, n. 2, p. 55–60, February 2002.

[Fensel 2000]FENSEL, D. The Semantic Web and its languages. *IEEE Intelligent Systems*, v. 15, n. 6, p. 67–73, November 2000.

[Fensel 2002]FENSEL, D. Ontology-based knowledge management. *IEEE Computer*, v. 35, n. 11, p. 56–59, November 2002.

[Fensel *et al.* 2003]FENSEL, D. *et al. Spinning the Semantic Web - bringing the World Wide Web to its full potential.* 13^{th} ed. [S.l.]: MIT Press, 2003. 392 p.

[Freitas, Stuckenschmidt e Noy 2005]FREITAS, F.; STUCKENSCHMIDT, H.; NOY, N. F. Ontology issues and applications - guest editor's introduction. *Journal of the Brazilian Computer Society*, p. 5–16, November 2005.

[Gasperin 2001]GASPERIN, C. V. *Extração automática de relações semânticas a partir de relações sintáticas.* Dissertação (Mestrado) — Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2001. 134 p.

[Giunchiglia e Shvaiko 2004]GIUNCHIGLIA, F.; SHVAIKO, P. Semantic matching. *The Knowledge Engineering Review*, v. 18, n. 3, p. 265–280, 2004.

[Giunchiglia, Yatskevich e Giunchiglia 2005]GIUNCHIGLIA, F.; YATSKEVICH, M.; GIUNCHIGLIA, E. Efficient semantic matching. In: *Proceedings of the European Semantic Web Conference - ESWC*, p. 272–289, 2005.

[Gruber 1993]GRUBER, T. R. A translation approach to portable ontology specifications. *Knowledge Acquisition*, v. 5, n. 2, p. 199–220, June 1993.

[Gruber 1995]GRUBER, T. R. Toward principles for the design of ontologies used for knowledge sharing. In: *International Journal of Human-Computer Studies*, v. 43, n. 5/6, p. 907–928, 1995.

[Gruninger e Lee 2002]GRUNINGER, M.; LEE, J. Ontology applications and design. *Communications of the ACM*, v. 45, n. 2, p. 39–41, February 2002.

[Guarino 1996]GUARINO, N. Understanding, building, and using ontologies. A commentary to "Using explicit ontologies in KBS development". In: *Proceedings of the* 10^{th} *Knowledge Aquisition for Knowledge-Based Systems Workshop*, 1996.

[Holsapple e Joshi 2002]HOLSAPPLE, C. W.; JOSHI, K. D. A collaborative approach to ontology design. *Communications of the ACM*, v. 45, n. 2, p. 42–47, February 2002.

[Houari e Far 2004]HOUARI, N.; FAR, B. H. Application of intelligent agent technology for knowledge management integration. In: *Proceedings of the* 3^{rd} *IEEE International Conference on Cognitive Informatics (ICCI 2004)*, p. 240–249, 2004.

[Jones 1986]JONES, K. S. *Synonymy and semantic classification*. Edinburgh: Edinburgh University Press, 1986. 285 p.

[Jurafsky e Martin 2000]JURAFSKY, D.; MARTIN, J. H. *Speech and language processing: an introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. 1^{st} ed. New Jersey: Prentice Hall, 2000. 960 p.

[Kalfoglou e Schorlemmer 2003]KALFOGLOU, Y.; SCHORLEMMER, M. Ontology mapping: the state of the art. *The Knowledge Engineering Review*, v. 18, n. 1, p. 1–31, January 2003.

[Lassila e Swick 2005]LASSILA, O.; SWICK, R. R. *RDF/XML syntax specification*. Acesso em: Novembro 2005. Disponível em: <http://www.w3.org/TR/rdf-syntax-grammar/>.

[Levenshtein 1966]LEVENSHTEIN, V. Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, v. 10, n. 8, p. 707–710, 1966.

[Lin 1998]LIN, D. An information-theoretic definition of similarity. In: *Proceedings of the* 15^{th} *International Conference on Machine Learning*, p. 296–304, 1998.

[Madhavan, Bernstein e Rahm 2001]MADHAVAN, J.; BERNSTEIN, P. A.; RAHM, E. Generic schema matching using Cupid. In: *Proceedings of the* 27^{th} *Very Large Data Bases (VLDB)*, p. 48–58, 2001. Disponível em: <http://research.microsoft.com/philbe/CupidVLDB01.pdf>.

[Maedche 2002]MAEDCHE, A. *Ontology learning for the Semantic Web*. 1^{st} ed. Boston: Kluwer Academic Publishers, 2002. 272 p.

[Maedche *et al.* 2002]MAEDCHE, A. *et al.* MAFRA - A MApping FRAmework for distributed ontologies. In: *Proceedings of the* 13^{th} *International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, Springer-Verlag, London, UK, 2002.

[Maedche e Staab 2002]MAEDCHE, A.; STAAB, S. Measuring similarity between ontologies. In: *Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW)*, 2002.

[Manning e Schütze 1999]MANNING, C. D.; SCHüTZE, H. *Foundations of Statistical Natural Language Processing*. 1^{st} ed. Cambridge, Massachusetts: MIT Press, 1999. 620 p.

[McGuinness e Harmelen 2005]MCGUINNESS, D. L.; HARMELEN, F. van. *OWL - Web Ontology Language overview*. Acesso em: Novembro 2005. Disponível em: <http://www.w3.org/TR/owl-features/>.

[Noy 2004]NOY, N. F. Semantic integration: a survey of ontology-based approaches. *SIGMOD Record*, v. 33, n. 4, p. 65–70, December 2004.

[Noy e McGuinness 2001]NOY, N. F.; MCGUINNESS, D. L. Ontology development 101: a guide to creating your first ontology. *Stanford Knowledge Systems Laboratory and Stanford Medical Informatics*, March 2001.

[Noy e Musen 2000]NOY, N. F.; MUSEN, M. A. Prompt: Algorithm and tool for automated ontology merging and alignment. *In: Proceedings of the National Conference on Artificial Intelligence*, 2000.

[Noy e Musen 2003]NOY, N. F.; MUSEN, M. A. The PROMPT suite: Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, v. 59, n. 6, p. 983–1024, 2003.

[Orengo e Huyck 2001]ORENGO, V. M.; HUYCK, C. A stemming algorithm for the Portuguese language. In: *Proceedings of the* 8^{th} *International Symposium on String Processing and Information Retrieval (SPIRE-2001)*, p. 186–193, 2001.

[Paice 1994]PAICE, C. D. An evaluation method for stemming algorithms. In: *ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 42–50, 1994.

[Pinto, Gómez-Pérez e Martins 1999]PINTO, H. S.; GóMEZ-PéREZ, A.; MARTINS, J. P. Some issues on ontology integration. In: *Proceedings of IJCAI99's Workshop on Ontologies and Problem Solving Methods: Lessons Learned and Future Trends*, v. 18, p. 7–1/7–12, August 1999.

[Porter 2006]PORTER, M. *Porter Stemming Algorithm*. Acesso em: Janeiro 2006. Disponível em: <http://www.tartarus.org/ martin/PorterStemmer/>.

[Serafini *et al.* 2003]SERAFINI, L. *et al.* An algorithm for matching contextualized schemas via SAT. *Instituto Trentino di Cultura*, January 2003.

[Shvaiko e Euzenat 2005]SHVAIKO, P.; EUZENAT, J. A survey of schema-based matching approaches. *Journal on Data Semantics IV*, v. 4, p. 146–171, 2005.

[Staab, Erdmann e Maedche 2000]STAAB, S.; ERDMANN, M.; MAEDCHE, S. D. A. An extensible approach for modeling ontologies in RDF(S). In: *Proceedings of First Workshop on the Semantic Web*, Lisbon, Portugal, April 2000.

[Stumme *et al.* 2000]STUMME *et al.* Fast computation of concept lattices using data mining techniques. In: *Proceedings of the* 7^{th} *International Workshop 'Knowledge Representation meets Databases' (KRDB 2000)*, p. 129–139, 2000.

[Stumme e Maedche 2001]STUMME, G.; MAEDCHE, A. FCA-Merge: bottom-up merging of ontologies. In: *Proceedings of the* 17^{th} *Intl. Conf. on Artificial Intelligence (IJCAI '01)*, p. 225–230, 2001.

[Stumme, Studer e Sure 2000]STUMME, G.; STUDER, R.; SURE, Y. Towards an order-theoretical foundation for maintaining and merging ontologies. In: *Tagungsband der Verbundtagung Wirtschaftsinformatik 2000*, p. 136–149, 2000.

[Uschold e King 1995]USCHOLD, M.; KING, M. Towards a methodology for building ontologies. In: *Proceedings of IJCAI95's Workshop on Basic Ontological Issues in Knowledge Sharing*, July 1995.

[Yule 1998]YULE, G. *The study of language.* 2^{nd} ed. Great Britain: Cambridge University Press, 1998. 294 p.

Apêndice A

Análise humana - Par 1

Tabela A.1: Análise humana (Lingüista) para o Par 1 de ontologias

	Termo O_1	Termo O_2
1	direito penal	direito penal
2	direito processual	direito processual
3	direito canônico	direito canônico
4	direito internacional	direito internacional privado
5	direito internacional	direito internacional público
6	direito administrativo	direito administrativo
7	direito aéreo	direito aéreo
8	direito civil	direito civil
9	direito comercial	direito comercial
10	direito constitucional	direito constitucional
11	direito do trabalho	direito do trabalho
12	filosofia do direito	filosofia do direito
13	direito militar	direito militar
14	filosofia do direito	teoria do direito
15	história do direito	fontes do direito
16	história do direito	filosofia do direito

Tabela A.2: Análise humana (Bacharel em Ciência da Computação) para o Par 1 de ontologias

	Termo O_1	Termo O_2
01	direito	direito
02	direito administrativo	direito administrativo
03	direito aéreo	direito aéreo
04	direito canônico	direito canônico
05	direito civil	direito civil
06	direito comercial	direito comercial
07	direito constitucional	direito constitucional
08	direito do trabalho	direito do trabalho
09	direito internacional	direito internacional privado
10	direito internacional	direito internacional público
11	direito militar	direito militar
12	direito penal	direito penal
13	direito processual	direito processual
14	filosofia do direito	filosofia do direito

Tabela A.3: Análise humana (Bacharel em Direito) para o Par 1 de ontologias

	Termo O_1	Termo O_2
1	direito administrativo	direito administrativo
2	direito aéreo	direito aéreo
3	direito canônico	direito canônico
4	direito civil	direito civil
5	direito comercial	direito comercial
6	direito constitucional	direito constitucional
7	direito do trabalho	direito do trabalho
8	direito internacional	direito internacional público
9	direito internacional	direito internacional privado
10	direito militar	direito militar
11	direito penal	direito penal
12	direito processual	direito processual
13	filosofia do direito	filosofia do direito

Apêndice B

Análise humana - Par 2

Tabela B.1: Análise humana (Lingüista) para o Par 2 de ontologias

	Termo O_1	Termo O_2
1	ato de comércio	ato de comércio
2	contrato comercial	contrato comercial
3	direito bancário	direito bancário
4	direito cambiário	direito cambiário
5	direito industrial	direito industrial
6	sociedade comercial	sociedade comercial
7	associação comercial	sociedade comercial
8	mercadoria	mercado de capitais
9	direito falimentar	concordata

Tabela B.2: Análise humana (Bacharel em Ciência da Computação) para o Par 2 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito comercial	direito comercial
3	ato de comércio	ato de comércio
4	contrato comercial	contrato comercial
5	direito bancário	direito bancário
6	direito cambiário	direito cambiário
7	direito industrial	direito industrial
8	sociedade comercial	sociedade comercial

Tabela B.3: Análise humana (Bacharel em Direito) para o Par 2 de ontologias

	Termo O_1	Termo O_2
1	direito comercial	direito comercial
2	ato de comércio	ato de comércio
3	ato de comércio	compra e venda
4	contrato comercial	contrato comercial
5	contrato comercial	compra e venda
6	direito bancário	direito bancário
7	direito cambiário	direito cambiário
8	direito industrial	direito industrial
9	direito empresarial	direito comercial
10	sociedade comercial	sociedade comercial

Apêndice C

Análise humana - Par 3

Tabela C.1: Análise humana (Lingüista) para o Par 3 de ontologias

	Termo O_1	Termo O_2
1	direito administrativo	direito administrativo
2	direito	direito
3	contencioso administrativo	contencioso administrativo
4	direito disciplinar	direito disciplinar
5	domínio público	domínio público
6	controle administrativo	poder administrativo
7	processo administrativo	processo administrativo
8	poder administrativo	poder administrativo
9	função pública	funcionário público
10	ato administrativo	ato administrativo
11	competência administrativa	competência administrativa
12	delegação de competência	competência administrativa
13	jurisdição administrativa	poder administrativo
14	jurisdição administrativa	competência administrativa
15	responsabilidade administrativa	competência administrativa

Tabela C.2: Análise humana (Bacharel em Ciência da Computação) para o Par 3 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito administrativo	direito administrativo
3	ato administrativo	ato administrativo
4	competência administrativa	competência administrativa
5	contencioso administrativo	contencioso administrativo
6	direito disciplinar	direito disciplinar
7	domínio público	domínio público
8	poder administrativo	poder administrativo
9	processo administrativo	processo administrativo

Tabela C.3: Análise humana (Bacharel em Direito) para o Par 3 de ontologias

	Termo O_1	Termo O_2
1	direito administrativo	direito administrativo
2	ato administrativo	ato administrativo
3	competência administrativa	competência administrativa
4	contencioso administrativo	contencioso administrativo
5	controle administrativo	poder administrativo
6	contencioso administrativo	processo administrativo
7	direito disciplinar	direito disciplinar
8	função pública	serviço público
9	jurisdição administrativa	competência administrativa
10	processo administrativo	processo administrativo
11	poder administrativo	poder administrativo

Apêndice D

Análise humana - Par 4

Tabela D.1: Análise humana (Lingüista) para o Par 4 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito eleitoral	direito eleitoral
3	eleições	eleição
4	eleição estadual	eleição estadual
5	eleição municipal	eleição municipal
6	eleição parlamentar	eleição parlamentar
7	eleição presidencial	eleição presidencial
8	eleição primária	eleição primária
9	justiça eleitoral	sistema eleitoral
10	partidos políticos	partido político
11	voto	voto
12	voto censitário	voto censitário
13	voto da mulher	voto da mulher
14	voto distrital	voto distrital
15	voto do analfabeto	voto do analfabeto
16	voto do menor	voto do menor
17	voto eletrônico	voto eletrônico
18	voto em branco	voto em branco
19	voto nulo	voto nulo
20	voto obrigatório	voto obrigatório
21	voto popular	voto popular
22	voto secreto	voto secreto

Tabela D.2: Análise humana (Bacharel em Ciência da Computação) para o Par 4 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito eleitoral	direito eleitoral
3	eleições	eleição
4	eleição estadual	eleição estadual
5	eleição municipal	eleição municipal
6	eleição parlamentar	eleição parlamentr
7	eleição presidencial	eleição presidencial
8	eleição primária	eleição primária
9	partidos políticos	partido político
10	voto	voto
11	voto censitário	voto censitário
12	voto da mulher	voto da mulher
13	voto distrital	voto distrital
14	voto do analfabeto	voto do analfabeto
15	voto do menor	voto do menor
16	voto eletrônico	voto eletrônico
17	voto em branco	voto em branco
18	voto nulo	voto nulo
19	voto obrigatório	voto obrigatório
20	voto popular	voto popular
21	voto secreto	voto secreto
22	votação	voto

Tabela D.3: Análise humana (Bacharel em Direito) para o Par 4 de ontologias

	Termo O_1	Termo O_2
1	direito eleitoral	direito eleitoral
2	eleição estadual	eleição estadual
3	eleição municipal	eleição municipal
4	eleição parlamentar	eleição parlamentar
5	eleição presidencial	eleição presidencial
6	eleição primária	eleição primária
7	partidos políticos	partido político
8	voto censitário	voto censitário
9	voto da mulher	voto da mulher
10	voto distrital	voto distrital
11	voto do analfabeto	voto do analfabeto
12	voto do menor	voto do menor
13	voto eletrônico	voto eletrônico
14	voto em branco	voto em branco
15	voto nulo	voto nulo
16	voto obrigatório	voto obrigatório
17	voto popular	voto popular
18	voto secreto	voto secreto

Apêndice E

Análise humana - Par 5

Tabela E.1: Análise humana (Lingüista) para o Par 5 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito internacional	direito internacional privado
3	direito internacional privado	direito internacional privado
4	direito internacional	direito internacional público
5	cláusula de nação mais favorecida	cláusula de nação mais favorecida
6	direito internacional público	direito internacional público
7	direito de guerra	direito de guerra
8	direito diplomático	direito diplomático
9	direito do mar	direito do mar
10	direito econômico internacional	direito econômico internacional
11	direito fluvial internacional	direito fluvial internacional
12	direito internacional penal	direito penal internacional
13	tratados internacionais	relações internacionais
14	justiça internacional	direito penal internacional
15	represália internacional	direito penal internacional

Tabela E.2: Análise humana (Bacharel em Ciência da Computação) para o Par 5 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito internacional	direito internacional privado
3	direito internacional	direito internacional público
4	direito internacional privado	direito internacional privado
5	direito econômico internacional	direito econômico internacional
6	cláusula de nação mais favorecida	cláusula de nação mais favorecida
7	zona econômica exclusiva	jurisdição internacional
8	direito de guerra	direito de guerra
9	direito diplomático	direito diplomático
10	direito do mar	direito do mar
11	direito fluvial internacional	direito fluvial internacional
12	direito internacional penal	direito penal internacional
13	justiça internacional	direito penal interncional
14	direito internacional público	direito internacional público

Tabela E.3: Análise humana (Bacharel em Direito) para o Par 5 de ontologias

	Termo O_1	Termo O_2
1	direito	direito
2	direito internacional privado	direito internacional privado
3	direito internacional	direito internacional público
4	direito internacional	direito internacional privado
5	direito econômico internacional	direito econômico internacional
6	direito de guerra	direito de guerra
7	direito diplomático	direito diplomático
8	direito do mar	direito do mar
9	direito fluvial internacional	direito fluvial internacional
10	justiça internacional	jurisdição internacional
11	tratados internacionais	relações internacionais

Apêndice F

Falsos positivos MT(CSC) do Par 4

Tabela F.1: Coeficientes de similaridade dos termos mapeados pelo MT utilizando CSC para o Par 4 de ontologias

	Termo O_1	Termo O_2	MT(CSC)
01	direito	direito constitucional	0.95
02	direito	direito eleitoral	1.0
03	direito	voto	0.73
04	direito eleitoral	direito	1.0
05	direito eleitoral	direito constitucional	0.95
06	direito eleitoral	voto	0.73
07	sucessão presidencial	eleição presidencial	0.75
08	voto	direito	0.73
09	voto	direito constitucional	0.7
10	voto	direito eleitoral	0.73

Apêndice G

Falsos positivos SiSe(CSC') do Par 4

Tabela G.1: Coeficientes de similaridade dos termos mapeados pelo SiSe utilizando CSC' para o Par 4 de ontologias

	Termo O_1	Termo O_2	MT(CSC)
01	direito	direito constitucional	0.95
02	direito	direito eleitoral	1.0
03	direito eleitoral	direito	1.0
04	direito eleitoral	direito constitucional	0.95
05	sucessão presidencial	eleição presidencial	0.8
06	partidos políticos	convenção partidária	0.75
07	partidos políticos	partido comunista	0.75
08	partidos políticos	partido conservador	0.75
09	partidos políticos	partido democrático	0.75
10	partidos políticos	partido liberal	0.75
11	partidos políticos	partido republicano	0.75
12	partidos políticos	partido socialista	0.75
13	partidos políticos	partido trabalhista	0.75
14	fidelidade partidária	partido político	0.75
15	fundo partidário	partido político	0.75
16	votação	voto censitário	0.75
17	votação	voto da mulher	0.75
18	votação	voto distrital	0.75
19	votação	voto do analfabeto	0.75
20	votação	voto do menor	0.75
21	votação	voto eletrônico	0.75
22	votação	voto em branco	0.75
23	votação	voto nulo	0.75
24	votação	voto obrigatório	0.75

25	votação	voto popular	0.75
26	votação	voto secreto	0.75

Apêndice H

Modelo documento de avaliação

PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO GRANDE DO SUL
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
FACULDADE DE INFORMÁTICA
GRUPO DE PROCESSAMENTO DA LINGUAGEM NATURAL

Informações do projeto

Aluno de Mestrado: Juliano Baldez de Freitas
Orientadora: Vera Lúcia Strube de Lima
Área: Processamento da Linguagem Natural / Similaridade Semântica entre ontologias
Página: http://www.inf.pucrs.br/ jfreitas
Email: jfreitas@inf.pucrs.br

Resumo

Este trabalho de mestrado tem como objetivo encontrar o mapeamento entre ontologias criadas individualmente. O mapeamento analisa duas estruturas ontológicas e compara, através de uma medida de similaridade, os termos destas. A medida de similaridade proposta neste trabalho é denominada SiSe (Similaridade Semântica), e analisa as relações hierárquicas em que os termos comparados estão em relação à ontologia aos quais os mesmos pertencem. Através desta medida podemos encontrar termos que possam ser mapeados de uma ontologia para outra através de coeficientes de similaridade retornado, entre os valores 0 e 1, onde o valor 1 representa uma combinação perfeita destes termos.

Instruções de avaliação

A avaliação da medida SiSe terá como parâmetro de comparação avaliações feitas por humanos. A avaliação humana (representada neste documento) tem como objetivo a avaliação e a indicação do mapeamento de termos similares entre duas ontologias de acordo com um julgamento humano. Nesta avaliação temos sete (7) pares de ontologias. Estas ontologias estão identificadas como Ontologia 1 (Vocabulário Controlado Básico da USP[1] (VCBUSP)) e Ontologia 2 (Vocabulário Básico do Senado Federal[2] (VCBS)). Para cada par de ontologias o avaliador humano deve indicar na tabela de mapeamento os termos considerados semanticamente similares (comparar pelo significado semântico do termo dentro da sua respectiva ontologia) colocando-os em suas respectivas colunas. A seguir temos um exemplo da indicação do mapeamento entre duas ontologias. Podem ocorrer casos em que, um termo em uma ontologia, possui mais de um mapeamento na outra ontologia.

Exemplo da avaliação humana

[1] http://143.107.73.99/Vocab/Sibix652.dll/MAC
[2] http://webthes.senado.gov.br/thes/

PAR X

Ontologia 1 (O_1)	Ontologia 2 (O_2)
01 direito	01 direito
02 criminalística	02 direito administrativo
03 direito administrativo	03 direito aéreo
04 direito aéreo	04 direito canônico
05 direito canônico	05 direito civil
06 direito civil	06 direito comercial
07 direito comercial	07 direito comparado
...	...

Por exemplo ao identificar a similaridade semântica entre o termo DIREITO CIVIL na Ontologia 1, e o termo DIREITO CIVIL na Ontologia 2, a indicação do mapeamento deve ser feita da seguinte maneira, na tabela de mapeamento. A similaridade entre os termos deve ser julgada pela semântica (sentido do termo e posição na hierarquia da ontologia) do mesmo e não somente pela similaridade lexical. Se a tabela de mapeamento dos pares de ontologias não for suficiente para os mapeamentos identificados, podem ser inseridas células nas mesmas.

TABELA DE MAPEAMENTO PAR X

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1	direito civil	direito civil
2		
3		
4		

Avaliação

Informações do avaliador

Nome avaliador:
Área de Atuação:
E-mail:
Telefone:

PAR 1

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 criminalística	02 direito administrativo
03 direito administrativo	03 direito aéreo
04 direito aéreo	04 direito canônico
05 direito canônico	05 direito civil
06 direito civil	06 direito comercial
07 direito comercial	07 direito comparado
08 direito constitucional	08 direito constitucional
09 direito do trabalho	09 direito do trabalho
10 direito econômico	10 direito internacional privado
11 direito eleitoral	11 direito internacional público
12 direito financeiro	12 direito judaico
13 direito internacional	13 direito militar
14 direito militar	14 direito penal
15 direito penal	15 direito privado
16 direito previdenciário	16 direito processual
17 direito processual	17 direito público
18 direito tributário	18 direito romano
19 direito urbanístico	19 filosofia do direito
20 filosofia do direito	20 fontes do direito
21 história do direito	21 jurisprudência
	22 sociologia jurídica
	23 teoria do direito

TABELA DE MAPEAMENTO PAR 1

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1		
2		
3		
4		

PAR 2

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito comercial	02 direito comercial
03 associação comercial	03 ato de comércio
04 ato de comércio	04 compra e venda
05 código de proteção e	05 concordata
defesa do consumidor	06 contrato comercial
06 contrato comercial	07 contrato de transporte
07 direito aeronáutico	08 dano
08 direito alfandegário	09 direito autoral
09 direito bancário	10 direito bancário
10 direito cambiário	11 direito cambiário
11 direito da informática	12 direito industrial
12 direito industrial	13 garantia
13 direito empresarial	14 mercado de capitais
14 direito falimentar	15 seguro
15 direito marítimo	16 sociedade comercial
16 mercadoria	
17 sociedade comercial	

TABELA DE MAPEAMENTO PAR 2

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1		
2		
3		
4		

PAR 3

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito administrativo	02 direito administrativo
03 administração pública	03 competência administrativa
04 ato administrativo	04 contencioso administrativo
05 competência administrativa	05 direito disciplinar
06 contencioso administrativo	06 direito financeiro
07 contrato administrativo	07 direito tributário
08 controle admisnistrativo	08 domínio público
09 delegação de competência	09 funcionário público
10 direito disciplinar	10 organização administrativa
11 domínio público	11 poder administrativo
12 função administrativa	12 poder de polícia
13 função pública	13 processo administrativo
14 imprevisibilidade	14 reversão
15 jurisdição administrativa	15 serviço público
16 moralidade administrativa	16 ato administrativo
17 processo administrativo	
18 poder administrativo	
19 responsabilidade administrativa	
20 tribunal administrativo	

TABELA DE MAPEAMENTO PAR 3

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1		
2		
3		
4		

PAR 4

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito eleitoral	02 direito constitucional
03 crime eleitoral	03 direito eleitoral
04 fraude eleitoral	04 campanha eleitoral
05 domicílio eleitoral	05 eleição
06 eleições	06 eleição direta
07 elegibilidade	07 eleição estadual
08 inelegibilidade	08 eleição indireta
09 eleição estadual	09 eleição municipal
10 eleição municipal	10 eleição parlamentar
11 eleição parlamentar	11 eleição presidencial
12 eleição presidencial	12 eleição primária
13 sucessão presidencial	13 partido político
14 eleição primária	14 convenção partidária
15 mandato eletivo	15 partido comunista
16 reeleição	16 partido conservador
17 reforma eleitoral	17 partido democrático
18 justiça eleitoral	18 partido liberal
19 tribunal eleitoral	19 partido republicano
20 tribunal regional federal	20 partido socialista
21 competência eleitoral	21 partido trabalhista
22 partidos políticos	22 sistema eleitoral
23 fidelidade partidária	23 voto
24 fundo partidário	24 voto censitário
25 sistema distrital	25 voto da mulher
26 voto	26 voto distrital
27 cédula eleitoral	27 voto do analfabeto
28 voto censitário	28 voto do menor
29 voto da mulher	29 voto eletrônico
30 voto distrital	30 voto em branco
31 voto do analfabeto	31 voto nulo
32 voto do menor	32 voto obrigatório
33 voto eletrônico	33 voto popular
34 voto em branco	34 voto secreto
35 voto nulo	
36 voto obrigatório	
37 voto popular	
38 voto secreto	
39 votação	

40 contagem de votos

TABELA DE MAPEAMENTO PAR 4

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1		
2		
3		
4		

PAR 5

Ontologia 1 (O_1) VCUSP	Ontologia 2 (O_2) VCBS
01 direito	01 direito
02 direito internacional	02 direito internacional privado
03 direito internacional privado	03 conflito de leis
04 direito comercial internacional	04 sentença estrangeira
05 trips	05 direito internacional público
06 direito econômico internacional	06 arbitragem internacional
07 coisas e bens de direito internacional	07 cláusula de nação mais favorecida
08 concorrência internacional	08 direito de guerra
09 cláusula de nação mais favorecida	09 direito diplomático
10 incoterms	10 direito do mar
11 zona econômica exclusiva	11 direito econômico internacional
12 competência internacional	12 direito fluvial internacional
13 direito iternacional público	13 direito internacional de desenvolvimento
14 direito comunitário	14 direitos humanos
15 direito de guerra	15 jurisdição internacional
16 direito diplomático	16 pessoa jurídica de direito internacional público
17 direito do mar	17 direito consular
18 direito fluvial internacional	18 direito penal internacional
19 direito nuclear	19 relações internacionais
20 direito internacional penal	
21 equilíbrio internacional	
22 estrangeiro	
23 jus gentium	
24 justiça internacional	
25 reconhecimento internacional	
26 represália internacional	
27 sociedade internacional	
28 tratados internacionais	

TABELA DE MAPEAMENTO PAR 5

	Termo Ontologia 1 (O_1)	Termo Ontologia 2 (O_2)
1		
2		
3		
4		