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

UMA PROPOSTA PARA DESCOBERTA AUTOMÁTICA DE RELAÇÕES NÃO-TAXONÔMICAS A PARTIR DE CORPUS EM LÍNGUA PORTUGUESA

VINICIUS HARTMANN FERREIRA

Dissertação apresentada como requisito parcial à obtenção do grau de Mestre em Ciência da Computação na Pontifícia Universidade Católica do Rio Grande do Sul

Orientadora: Prof. Dra. Renata Vieira
Co-Orientadora: Dra. Lucelene Lopes

Porto Alegre
2012

F383p Ferreira, Vinicius Hartmann
Uma proposta para descoberta automática de relações não-taxonômicas a partir de corpus em língua portuguesa / Vinicius Hartmann Ferreira. – Porto Alegre, 2012.
86 f.

Diss. (Mestrado) – Fac. de Informática, PUCRS.
Orientador: Profª. Drª. Renata Vieira.
Co-Orientador: Drª. Lucelene Lopes.

1. Informática. 2. Ontologia. 3. Processamento da Linguagem Natural. I. Vieira, Renata. II. Lopes, Lucelene. III. Título.

CDD 006.35

Ficha Catalográfica elaborada pelo
Setor de Tratamento da Informação da BC-PUCRS

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 "Uma Proposta para Descoberta Automática de Relações Não-Taxonômicas a partir de Corpus em Língua Portuguesa" apresentada por Vinicius Hartmann Ferreira como parte dos requisitos para obtenção do grau de Mestre em Ciência da Computação, Inteligência Computacional, aprovada em 11/12/2012 pela Comissão Examinadora:

Profa. Dra. Renata Vieira – Orientador — PPGCC/PUCRS

Prof. Dr. Paulo Henrique Lemelle Fernandes – PPGCC/PUCRS

Dra. Lucelene Lopes – Pesquisadora/FACIN

Prof. Dr. Sandro José Rigo – UNISINOS

Homologada em 28/02/2013, conforme Ata No. 03/13 pela Comissão Coordenadora.

Prof. Dr. Paulo Henrique Lemelle Fernandes
Coordenador

PUCRS
Campus Central
Av. Ipiranga, 6681 – P32– sala 507 – CEP: 90619-900
Fone: (51) 3320-3611 – Fax (51) 3320–3621
E-mail: ppgcc@pucrs.br
www.pucrs.br/facin/pos

Para minha família.

AGRADECIMENTOS

Agradeço, em primeiro lugar, a Deus. Sem a Sua sustentação a realização desse trabalho não seria possível.

Aos meus pais, Miriam e Edevaldes, por sempre acreditarem em mim sem jamais duvidar de minhas capacidades.

Aos meus tios, Artur, Cledi, Elaine e Iolanda, pelo apoio e carinho sempre presentes.

A minha namorada, Elisângela, pelo incentivo, compreensão e companheirismo.

A minha orientadora, Dra. Renata Vieira, pela confiança que depositou em mim.

A minha co-orientadora, Dra. Lucelene Lopes, pela atenção e pelo conhecimento compartilhado.

Aos colegas e direção do IFRS, pelo apoio e compreensão.

Ao meu amigo, André, pelos conselhos e ajuda em todos momentos.

Ao meu amigo, Rafael, pelos conselhos e motivação.

Ao projeto PaleoProspec, pelo suporte fornecido. Também agradeço aos integrantes do projeto, companheiros de laboratório, pela amizade e auxílio.

Agradeço também aqueles que não acreditaram em mim, sem eles a realização desse trabalho também não seria possível.

UMA PROPOSTA PARA DESCOBERTA AUTOMÁTICA DE RELAÇÕES NÃO-TAXONÔMICAS A PARTIR DE CORPUS EM LÍNGUA PORTUGUESA

RESUMO

A construção de ontologias é um processo complexo que compreende etapas como a extração de conceitos de domínio, bem como a extração de relações taxonômicas e não-taxonômicas entre esses conceitos. A etapa de extração de relações não-taxonômicas é a mais negligenciada, especialmente para textos na língua portuguesa. Essa dissertação apresenta uma proposta de extração de relações não-taxonômicas a partir de textos em língua portuguesa (*corpora*). Esses textos são representados por uma lista de conceitos e informações contextuais automaticamente extraídos pela ferramenta ExATOlp. Uma aplicação do processo proposto foi realizada com *corpora* de cinco domínios e uma análise sobre a relevância dos conceitos, a especifidade das relações e a aplicação das relações extraídas foi realizada. Através dessa análise o processo proposto mostrou-se relevante, sendo considerado a principal contribuição dessa dissertação. Adicionalmente, uma ferramenta para visualização das relações não-taxonômicas extraídas, útil para diversas aplicações linguísticas, também é proposta.

Palavras-chave: Ontologias, Relações não-taxonômicas, Extração de relações de corpus, ExATOlp.

A PROPOSAL FOR AUTOMATIC DISCOVERY OF NON-TAXONOMIC RELATIONS FROM CORPUS IN PORTUGUESE

ABSTRACT

The construction of ontologies is a complex process that includes steps such as extraction of domain concepts, as well as the extraction of taxonomic and non-taxonomic relations between these concepts. The step of extracting non-taxonomic relations is the most neglected, specially for texts in portuguese. This dissertation presents a proposal for extracting non-taxonomic relations from texts in portuguese (*corpora*). These texts are represented by a list of domain concepts and contextual informations extracted by the tool ExATOlp. An application of the proposed process was performed with *corpora* of five domains and analysis on the relevance of the concepts, the specificity of relations and relations extracted application was made. Through this analysis, the proposed process seemed to be relevant and is considered the main contribution of this dissertation. Additionally, a tool for visualizing the extracted non-taxonomic relations, useful for various linguistic applications, is also proposed.

Keywords Ontologies, Non-taxonomic relations, Extraction of relations from corpus, ExATOlp

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE EQUAÇÕES

LISTA DE ABREVIATURAS

AO	Aprendizagem de Ontologias
SGBD	Sistema Gerenciador de Banco de Dados
PLN	Processamento de Linguagem Natural

SUMÁRIO

1 INTRODUÇÃO

Ontologias são uma das principais formas de representação de conhecimento de um domínio [1, 5, 13, 15]. De acordo com Cimiano, Volker e Studer [8], ontologias são compostas por um conjunto de conceitos referentes a um domínio específico e um conjunto de relações definidas entre os conceitos. Ontologias são de grande importância para sistemas de gestão de conhecimento, para a Web-Semântica e para a área de Processamento de Línguagem Natural (PLN) [8, 16, 21, 29, 31].

A construção manual de ontologias depende de engenheiros de ontologias auxiliados por um ou mais especialistas de domínio. Esta tarefa torna-se custosa e tediosa devido o tamanho e complexidade do domínio que está sendo modelado. Sendo assim, surge a necessidade de automatizar este processo [2, 16, 29].

Aprendizagem de Ontologias (AO) emprega técnicas de PLN, Mineração de Dados e Aprendizagem de Máquina para auxiliar o processo de construção de ontologias, propondo técnicas que tornem esse processo automático ou semi-automático [2, 16, 21, 29, 35]. A AO pode utilizar como fonte dados não-estruturados (textos em língua natural), semi-estruturados (dicionários) ou estruturados (*schemas* de bases de dados) [31].

Para Maedche e Staab [21], documentos não-estruturados são a maior fonte de conhecimento disponível. Neste contexto, podem ser citados muitos trabalhos que propõe processos de AO a partir de textos [2, 5, 8, 16, 21, 31, 35].

O processo para construir ontologias a partir de textos descrito por Buitelaar, Cimiano e Magnini [5] e sintetizado por Maedche e Staab [21], contempla três etapas básicas: (i) extração de conceitos de domínio; (ii) extração de taxonomia; e (iii) extração de relações não-taxonômicas.

A extração de informação em textos pode ocorrer de diversas formas, sendo uma das mais importantes à busca de conceitos em *corpora* de domínio. *Corpora* é o plural de *corpus,* que pode ser definido como um conjunto de textos sobre um domínio específico [19]. Nesse contexto, muitas iniciativas de extração de conceitos em *corpus* de domínio como as de Lopes [19], Pantel e Lin [26], Chung [7], Milios *et al.* [22], Drouin [9], Park *et al.* [27] e Kim *et al*. [18] partem de um processo básico de extrair os termos e estimar a sua relevância a fim de identificar os conceitos.

Como produto final do processo de extração apresentado nos trabalhos de Lopes [19], Milios *et al.* [22] e Drouin [9] gera-se um recurso linguístico composto por um conjunto dos conceitos mais relevantes do domínio aos quais são associadas informações contextuais. Entre as informações armazenadas, estão a forma canônica e a etiqueta sintática do conceito (substantivo, adjetivo, etc), a função gramatical do conceito na oração (sujeito, objeto, etc) e o predicado ao qual o conceito exerce sua função gramatical (verbo relacionado).

1.1 Motivação

De acordo com Maedech e Staab [21], existem muitas propostas de AO para facilitar a automatização da descoberta de conhecimento em textos [6, 10, 15, 23], porém estas propostas se focam apenas na extração de conceitos e na parte taxonômica das ontologias. A maior parte das propostas coleta conceitos relevantes de um domínio e os agrupa em uma hierarquia utilizando métodos linguísticos e estatísticos.

No processo de AO a fase de extração de relações não-taxonômicas tem sido reconhecida como a mais complexa [21, 33] e também a mais negligenciada [29, 33]. Por relação não-taxonômica entende-se a relação entre conceitos, geralmente através de um verbo, que não se baseia em hierarquia. Um exemplo deste tipo de relação pode ser visto no campo de Direito, no qual encontra-se a relação “representa” entre os conceitos “Advogado” e “Cliente” [31].

Na literatura podem ser encontradas propostas de extração de relações não-taxonômicas de textos como as de Sanchéz e Moreno [29], Villaverde *et al.* [33], Maedech e Staab [21], Serra e Girardi [31], de Schutz e Buitelaar [30] e de Weichselbraun *et al.* [35]. Todas essas propostas utilizam como fonte de dados textos em língua natural e nenhuma delas pode ser aplicada na língua portuguesa. E embora existam propostas para recuperação de informações em língua portugesa

produzidas e em andamento pelo Grupo de Processamento de Linguagem Natural da PUCR-RS, nenhuma delas se foca na extração de relações entre conceito sujeito e conceito objeto identificadas por um verbo.

Verifica-se também que nenhuma das propostas encontradas na literatura faz uso de informações contextuais relacionadas aos conceitos. Como as relações não-taxonômicas são indicadas por verbos, através das informações contextuais torna-se simples verificar a relação de um conceito com um verbo, sem a necessidade de um novo processo de recuperação de informações.

A ferramenta ExATOlp [20], operacionaliza o processo de extração de conceitos proposto por Lopes [19]. Uma das saídas da ferramenta, é uma tabela com termos candidatos a conceitos de um domínio e suas informações contextuais e outra saída é uma tabela com os conceitos de domínio e seu índice de relevância (*tf-dcf* [36]). Verificou-se então que essas informações contextuais extraídas para os conceitos proveêm informações necessárias para a identificação de relações não-taxonômicas.

Com isto, este trabalho tem como motivação o fato de: (i) a extração de relações não-taxonômicas ser uma área pouco explorada; (ii) não existir propostas de extração para a língua portuguesa; e (iii) nenhuma das propostas ter como fonte as informações contextuais relacionadas aos conceitos.

1.2 Objetivos

O objetivo geral do trabalho é propor um processo para automatizar a descoberta de relações não-taxonômicas para ontologias construídas a partir de *corpus* em língua portuguesa. Para que este objetivo seja alcançado, foram definidos os seguintes objetivos específicos:

I. Pesquisar fontes bibliográficas para levantamento de informações necessárias para a extração de relações não-taxonômicas;

II. Implementar um sistema computacional que operacionalize o processo proposto neste trabalho; e

III. Avaliar as relações não-taxonômicas extraídas através do processo proposto neste trabalho.

Para que os objetivos traçados nesta proposta sejam concluídos, na seção seguinte será descrita a metodologia que será utilizada.

1.3 Metodologia

Observou-se nos trabalhos similares que antes de iniciar a etapa de extração de relações não-taxonômicas, os processos propostos se preocupam com a extração de conceitos dos *corpus* de domínio. Após esta etapa, é feita a busca de conceitos que se relacionam através de um verbo nos textos não-estruturados. Porém, nem sempre estes conceitos utilizados são os mais relevantes de um domínio.

Sendo assim, ao analisar propostas de AO a partir de textos, pode-se citar os trabalhos de Lopes [19], Milios *et al.* [22] e Drouin [9], que além de extraírem os conceitos mais relevantes de um domínio também registram informações contextuais dos conceitos. Através do contexto de aplicação de um conceito é possível identificar informações como a função sintática do conceito em uma sentença, assim como o verbo com o qual ele está relacionado. Dos trabalhos citados, apenas o proposto por Lopes [19] tem como foco textos da língua portuguesa, que são o foco também deste trabalho.

O diferencial do processo que essa dissertação propõe em relação aos trabalhos similares, além de ter como foco textos da língua portuguesa, está na fonte de onde as relações serão extraídas. Como uma sentença na língua portuguesa tem como esrutura sintática básica "Sujeito + Verbo + Objeto", e verificou-se que a principal abordagem para descoberta de relações não-taxonômicas se dá pela relação entre conceitos através de um verbo, concluiu-se que as informações contextuais provêem recursos necessários para extrair relações não-taxonômicas. Sendo assim, o processo proposto terá como fonte uma lista dos conceitos mais relevantes de um domínio e suas informações contextuais geradas através do software ExATOlp, que implementa todos passos do processo proposto por Lopes [19].

Na lista de conceitos e informações contextuais gerada pelo ExATOlp, é possível identificar os verbos (em sua forma canônica) aos quais os conceitos se relacionam. Com isso, dois conceitos que se relacionam através de um mesmo verbo irão compor uma tripla definida por <*Conceito 1, verbo, Conceito 2*>. Também é possível identificar a função gramatical dos conceitos da tripla, caso

um deles seja identificado como sujeito e outro como objeto de uma sentença, dessa forma, a tripla é adicionada a lista de relações não-taxonômicas do domínio.

As relações não-taxonômicas extraídas através do processo proposto nessa dissertação, foram verificadas de três formas de análise: (i) análise da especificidade das relações (verbos) extraídas, através da comparação com verbos encontrados no *corpus* do jornal Diário Gaúcho; (ii) análise dos conceitos de domínio identificados como sujeitos ou objetos; e (iii) análise das relações não-taxonômicas extraídas do ponto de vista de um especialista no contexto de análise de papéis semânticos.

1.4 Organização

A Dissertação está organizada da seguinte forma:

- O Capítulo 2 apresenta a revisão bibliográfica sobre os conceitos fundamentais para a compreensão deste trabalho, sendo eles Ontologias, Aprendizagem de Ontologias e Relações Não-Taxonômicas. Também é apresentada neste Capítulo a revisão e análise dos trabalhos similares ao proposto nessa dissertação.

- O Capítulo 3 apresenta o processo proposto nessa dissertação, detalhando cada uma de suas etapas. Nesse Capítulo também é descrita e apresentada a ferramenta para visualização de relações não-taxonômicas desenvolvida.

- O Capítulo 4 relata a execução do processo proposto através de aplicações e seus seus resultados. Também neste Capítulo é descrita a metodologia utilizada para analisar as relações não-taxonômicas obtidas, assim como os resultados e a discussão da análise.

- O Capítulo 5 apresenta as conclusões obtidas com a realização dessa dissertação, além de propôr trabalhos futuros.

2 CENÁRIO E CONTEXTUALIZAÇÃO

Neste capítulo serão apresentados conceitos fundamentais para a compreensão desse trabalho. Na seção 2.1 será apresentada a definição para ontologias, sua importância e contribuição para a Web Semântica. Na seção 2.2 será descrito o processo de Aprendizagem de Ontologias, que auxilia na construção automática ou semi-automática de ontologias. Na seção 2.3 será apresentado o conceito de relações não-taxonômicas, um tema pouco abordado por propostas de Aprendizagem de Ontologias. Na seção 2.4 serão apresentados trabalhos de extração de relações não-taxonômicas a partir de textos, similares a que essa dissertação se propõe a fazer.

2.1 Ontologias

Para a filosofia, ontologia é a teoria sobre a natureza da existência, sobre quais tipos de coisas existem. Porém, para a Ciência da Computação, o conceito não é tão genérico. A partir das pesquisas realizadas nas áreas de Inteligência Artificial o termo ontologia foi adaptado e refere-se a um documento ou arquivo que define formalmente a relação entre conceitos de um determinado domínio [1, 13].

Na definição de Gruber [13], ontologia é uma especificação formal e explícita de uma conceitualização compartilhada por um domínio de interesse. Esta definição é acrescida pela de Swartout [32], na qual uma ontologia é um conjunto de termos hierarquicamente estruturados com o objetivo de descrever um domínio que pode ser usado como fundamento de uma base de conhecimento.

Uma ontologia é composta de um conjunto de conceitos ou classes de interresse referentes a um domínio específico e um conjunto de relações definidas entre estes conceitos. Estas relações podem ser classificadas em: (i) taxonômicas, que é a classificação hierárquica dos conceitos e; (ii) não-taxonômicas, que definem relações entre os conceitos sem agrupá-las em hierarquias [8, 19].

Para Brewster *et al.* [4], as ontologias são a forma fundamental de representação de conhecimento nos sistemas de Inteligência Artificial contemporâneos. Conforme Berners-Lee, Hendler e Lassila [1], as ontologias podem melhorar as funcionalidades da Web de muitas formas. Uma delas, por exemplo, é permitindo que buscadores Web encontrem material relacionado apenas

ao conceito pesquisado ao invés de encontrarem páginas associadas a palavras ambíguas. Pode-se também destacar o papel fundamental das ontologias em sistemas de gestão de conhecimento, na Web-Semântica, em sistemas de *e-commerce* e na área de PLN [5, 16, 29, 30, 31].

De acordo com Noy e McGuinness [25], uma ontologia define um vocabulário comum para pesquisadores que necessitam compartilhar ou apresentar informações sobre um determinado domínio. Dentro deste contexto, citam-se como razões para o desenvolvimento de uma ontologia a necessidade de permitir reuso de conhecimento, apresentar conceitos fundamentais de um domínio para um leigo, analisar o conhecimento de um domínio e separar o conhecimento de um domínio do conhecimento operacional.

A construção de ontologias de domínio depende de especialistas de domínio e de engenheiros do conhecimento, que são muitas vezes sobrecarregados pelo grande tamanho, complexidade e dinamicidade de um determinado domínio. Consequentemente, a construção manual de ontologias pode tornar-se um processo lento e custoso que necessita de métodos que auxiliem a sua construção [3, 29].

2.2 Aprendizagem de Ontologias

Ao processo de construir ontologias de forma semi-automática ou automática dá-se o nome de "Aprendizagem de Ontologia" (AO) [2, 16, 21, 29, 35]. Exitem dois aspectos fundamentais que devem ser observados em AO. O primeiro é a disponibilidade de conhecimento prévio, que pode ser na forma de uma ontologia para ser extendida. O segundo são as fontes de dados de onde será extraído conhecimento, que podem ser: (i) fontes não-estruturadas, como documentos em língua natural; (ii) fontes semi-estruturadas, como dicionários; e (iii) fontes estruturadas, como *schemas* de bases de dados [31].

Recentemente, a AO a partir de textos tem sido sugerida como uma tecnologia promissora, pois de acordo com Bieman [2] textos em língua natural são a maior fonte de conhecimento disponível. Este processo baseia-se na combinação de análise de textos, mineração de dados e modelagem de conhecimento [16].

O processo para construir ontologias a partir de textos é constituído de cinco etapas: (i) extração de termos candidados a conceitos de um domínio; (ii) identificação de sinônimos entre os

termos candidadtos a conceito; (iii) identificação hierárquica entre os conceitos; (iv) identificação das relações entre os conceitos; e (v) população da ontologia [5] (Figura 1).

Figura 1. Etapas de Aprendizagem de Ontologias

Fonte: Lopes [19]

O processo deve ser executado na ordem apresentada pela organização das camadas. De acordo com Buitelaar, Cimiano e Magnini [5], a primeira etapa é de suma importância para a qualidade da ontologia construída. Pois se a extração de termos candidatos a conceitos ocorrer de forma deficiente, nenhuma das outras etapas poderá compensar essa deficiência.

A extração de informação em textos pode ocorrer de diversas formas, sendo uma das mais importantes a busca de conceitos em *corpora* de domínio. *Corpora* é o plural de *corpus*, que pode ser definido como um conjunto de textos sobre um domínio específico [19]. Nesse contexto, as iniciativas de extração de conceitos em *corpus* de domínio propostos em trabalhos como os de Lopes [19], Pantel e Lin [26], Chung [7], Milios *et al*. [22], Drouin [9], Park *et al.* [27] , Kim *et al.* [18] e Brewster *et al.* [4] partem de um processo básico de identificar os termos e estimar a sua relevância a fim de identificar os conceitos.

Dentre as propostas de extração automática de conceitos, a única que se aplica a língua portuguesa é a apresenta por Lopes [19]. O processo proposto por Lopes [19] pode ser dividido em quatro etapas: (i) extração de termos e conceitos; (ii) classificação de termos; (iii) identificação de conceitos; e (iv) aplicações (Figura 2).

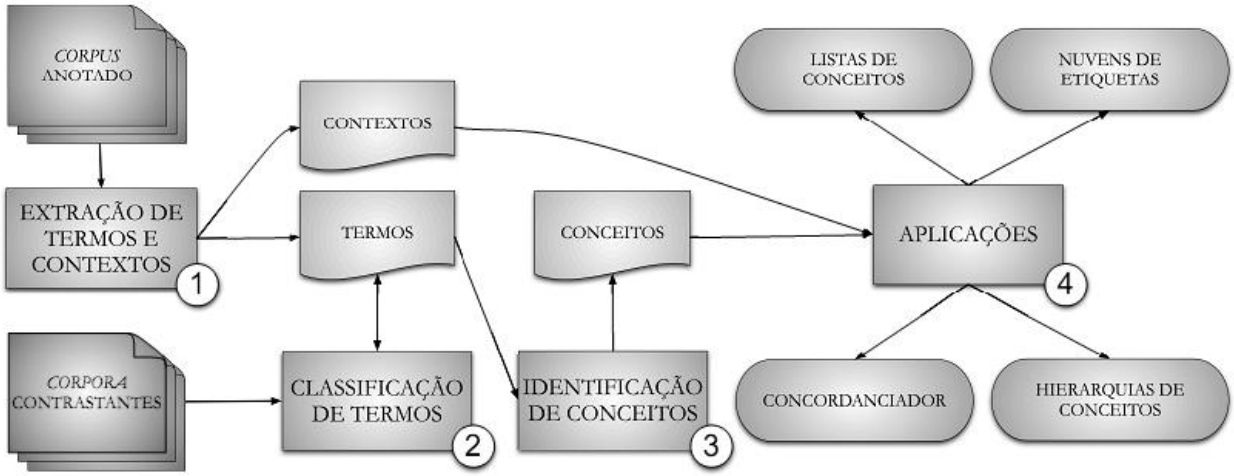

Figura 2. Processo de extração automática de conceitos

Fonte: Lopes [19]

A etapa de extração de conceitos corresponde a um processo linguístico que recebe um *corpus* de domínio anotado e identificam-se os termos candidatos a conceitos deste domínio. Também são extraídas informações sobre como estes conceitos foram utilizados no corpus (contexto) e o número das ocorrências em que o termo foi encontrado em cada uma das suas situações de uso no *corpus*.

Na etapa de classificação os termos são ordenados de acordo com a sua relevância através de um processo estatístico que leva em conta o *corpus* de domínio e um conjunto de *corpora* utilizados como contraste desse domínio. A etapa de identificação dos conceitos consiste em receber os termos classificados de acordo com sua relevância e aplicar um ponto de corte para definir quantos termos devem ser considerados conceitos do domínio. A última etapa do processo consiste em utilizar os conceitos e as informações contextuais em aplicações linguísticas.

Nos trabalhos encontrados na literatura, como produto dos processos de extração de conceitos em textos é gerada uma lista de conceitos. Além disto, os trabalhos de Lopes [19], Milios *et al.* [22] e Drouin [9] também registram informações contextuais dos conceitos encontrados, porém em nenhum destes trabalhos estas informações são apresentadas de forma tão declarada quanto no trabalho de Lopes [19]. Estas informações contextuais registradas junto aos conceitos podem ser visualizadas na Tabela 1.

Tabela 1. Conceitos e informações conceituais extraídas

Fonte: Lopes [19]

Identificador	Descrição
1	Conceito na sua forma original
2	Conceito na sua forma canônica
3	Número de palavras que compõem o conceito (1 para unigramas, 2 para bigramas, 3 para trigramas, etc)
4	Função gramatical do termo na oração (sujeito, objeto, etc)
5	Palavra indicada como núcleo do conceito na sua forma canônica
6	Etiqueta sintática do núcleo (substantivo, adjetivo, etc)
7	Etiqueta morfológica do núcleo
8	Etiqueta(s) semântica(s) do núcleo (uma estivamativa feita pelo parser)
9	Extras – etiquetas semânticas extras que o parser adiciona
10 e 11	Posição ocupada pelo conceito na frase (onde situam-se as palavras que compoem o conceito)
12	Número total de palavras na frase
13	Identificador da frase de onde o conceito foi extraído
14	Identificador do documento de onde o conceito foi extraído
15	Identificador do *corpus* de onde o conceito foi extraído
16	Predicado (verbo) ao qual o conceito exerce sua função gramatical na forma original
17	Predicado (verbo) ao qual o conceito exerce sua função gramatical na forma canônica
18	Etiquetas morfológicas dos verbos
19 e 20	Posição ocupada pelo predicado ao qual o conceito exerce sua função gramatical na frase

De acordo com Maedche e Staab [21], o processo de construção de ontologias contempla três etapas básicas: (i) extração de conceitos de domínio; (ii) extração de taxonomia; e (iii) extração de relações não-taxonômicas. Existem muitas propostas de AO que propõem métodos para facilitar a automatização da descoberta de conhecimento em textos [4, 6, 10, 15, 23], porém estas propostas se focam apenas na extração de conceitos e na parte taxonômica das ontologias.

A maior parte das propostas coleta conceitos relevantes de um domínio e os agrupa em uma hierarquia (taxonomia) utilizando métodos linguísticos e estatísticos [21]. De acordo com Sanchéz e Moreno [29], no processo de AO a fase de extração de relações não-taxonômicas tem sido reconhecida como a mais complexa [21, 33] e também a mais negligenciada [29, 33].

2.3 Relações Não-Taxonômicas

Diferente das relações taxonômicas, que contribuem na estruturação de um domínio e classificação de conceitos, as relações não-taxonômicas não estão relacionadas a hierarquia. Este tipo de relação acrescenta informações aos conceitos já encontrados, identificando os relacionamentos entre eles [31].

Identificar as relações não-taxonômicas é essencial para expressar as propriedades tanto de classes quanto de entidades de um domínio específico [8], representando as ações ou eventos que ocorrem entre os conceitos [29]. Podem ser citados como exemplos de relações não-taxonômicas no campo de Direito, a relação "representa" entre os conceitos "Advogado" e "Cliente" [31] e no campo dos Esportes a relação "chuta" entre os conceitos "Jogador" e "Bola" [30].

De acordo com Serra e Girardi [31], relações não-taxonômicas podem ser classificadas como independentes ou dependentes de domínio. As relações independentes de domínio podem ser divididas em: (i) agregação, identificadas por relações "todo-parte"; e (ii) propriedade, identificadas por relações de posse ou composição. As relações dependentes de domínio são identificadas por termos específicos de um domínio. Um exemplo de cada tipo de relação não-taxonômica é apresentado na Tabela 2.

Tabela 2. Classificação de relações não-taxonômicas.

Classificação	**Sub-Categoria**	**Exemplo**
Dependente de domínio	-	O advogado **representa** o cliente no julgamento.
Independente de domínio	Agregação	Um carro típico **tem** quatro rodas.
	Propriedade	Os pais irão aguardar a *decisão* **do** *tribunal*.

A classificação das relações não-taxonômicas em dependentes ou independentes de domínio não se aplica ao escopo deste trabalho. Embora seja possível classificá-las também na língua

portuguesa, não é um processo tão trivial quanto na língua inglesa. Isto ocorre pelo fato de que na língua inglesa as relações independentes de domínio são identificadas pelo apóstrofo, que indica a versão contraída do verbo *have* (ter), situação que não ocorre na língua portuguesa. Um exemplo de relação não-taxonômica independente de domínio identificada pelo apóstrofo é a frase "*Father and mother will wait for the court's decision*" (O pai e a mãe aguardarão a decisão do tribunal). Neste exemplo é identificada a relação não-taxonômica entre os conceitos tribunal (*court*) e decisão (*decision*).

O papel dos verbos como elemento de conexão central entre conceitos é inegável. Eles são responsáveis por especificar a interação entre os participantes de uma ação ou evento, expressando a relação entre eles. Devido a isto os verbos tem sido muito utilizados para definir relações não-taxonômicas [17, 21, 29, 30].

A descoberta de relações não-taxonômicas entre conceitos de uma ontologia pode ser dividida em duas atividades: (i) descobrir quais conceitos se relacionam; e (ii) nomear a relação [21, 29, 33]. O processo de identificar e nomear relações não-taxonômicas de forma manual é uma atividade pouco trivial, pois podem ser descobertas diversas relações diferentes entre os mesmos conceitos e quando existirem relações semelhantes é necessário definir a relação mais utilizada no domínio específico [29].

2.4 Trabalhos Similares

Com o objetivo de auxiliar a descoberta de relações não-taxonômicas, os trabalhos de Sanchéz e Moreno [29], Villaverde et al. [33], Maedech e Staab [21], Serra e Girardi [31], Schutz e Buitelaar [30] e Weichselbraun *et al.* [35] apresentam propostas para automatizar este processo.

O trabalho de Sanchéz e Moreno [29] apresenta uma proposta para descoberta de relações não-taxonômicas para construção de uma ontologia que tem como fonte a Web. O método proposto permite descobrir verbos relevantes para um domínio, e estes são utilizados como base de conhecimento para extrair e nomear relações não-taxonômicas de forma automática e sem supervisão (Figura 3).

Diferente da extração proposta nessa dissertação, que se aplica à língua portuguesa, o trabalho de Sanchéz e Moreno [29] restringe-se a língua inglesa. Porém, em ambos os trabalhos

considera-se o verbo como elemento central de conexão entre conceitos no processo de extração de relações não-taxonômicas.

Como a aprendizagem de ontologias é uma tarefa complexa, o processo de aprendizagem de Sanchéz e Moreno [29] foi projetado com características iterativas e incrementais, dividindo-se em duas etapas básicas: (i) aprendizagem de taxonomia; e (ii) aprendizagem de relações não-taxonômicas. Com isto, o conhecimento adquirido na primeira etapa pode ser utilizado para aperfeiçoar a segunda etapa, permitindo por exemplo a construção de queries de pesquisas mais restritivas.

O processo de Sanchéz e Moreno [29] começa através da busca de uma palavra-chave do domínio na Web. Inicialmente, somente consultas básicas usando padrões independentes de domínio para descoberta de hipônimos são utilizadas com o objetivo de recuperar *corpus* de documentos em sistemas de buscas na Web. Através desse processo é possível extrair o conhecimento mais diretamente relacionado ao domínio da pesquisa e compôr uma taxonomia inicial, como por exemplo tipos de câncer para o domínio câncer.

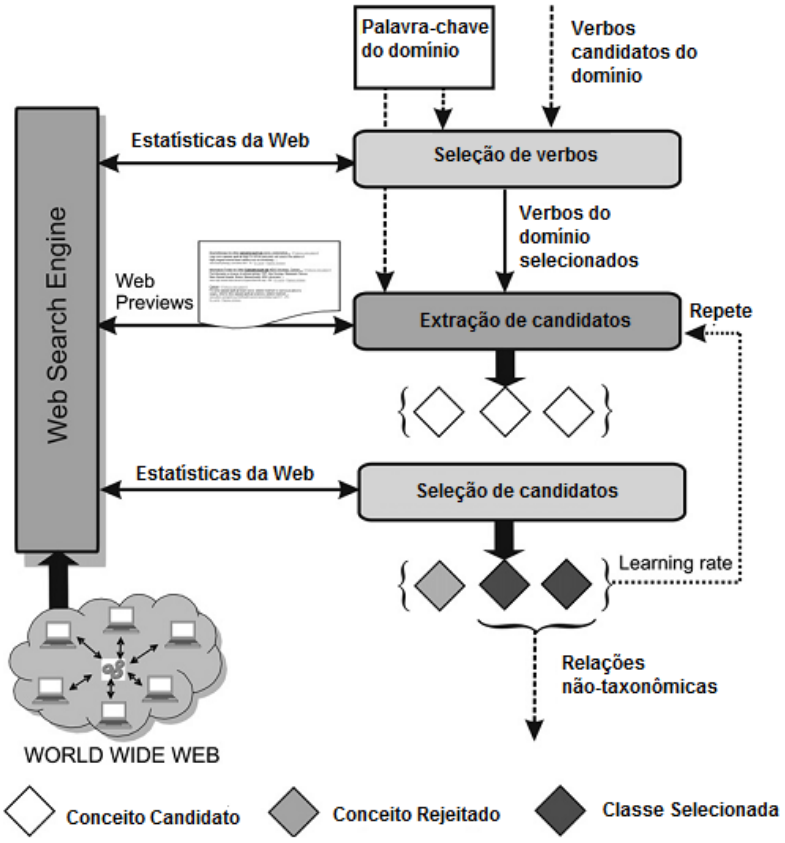

Figura 3. Processo para extração de relações não-taxonômicas na Web

Fonte: Adaptado de Sanchéz e Moreno [29]

Ainda nesse processo é executada uma tarefa para a descoberta de entidades relacionadas ao domínio, como por exemplo, organizações de saúde relacionadas ao domínio câncer. O produto final dessa tarefa é uma taxonomia com termos e verbos encontrados na mesma sentença que a palavra-chave de domínio usada nas pesquisas.

O processo de Sanchéz e Moreno [29] utiliza o conjunto de verbos encontrados na fase de aprendizagem de taxonomia e a palavra-chave do domínio como base de conhecimento para a extração de relações não-taxonômicas. De acordo com Sanchéz e Moreno [29], eles são utilizados como o ponto inicial para a construção de queries de pesquisa em sistemas de busca na Web.

A segunda etapa, aprendizagem de relações não-taxonômicas, divide-se em duas tarefas: (i) descoberta e seleção dos verbos utilizados para expressar relações não-taxonômicas e; (ii)

descoberta e seleção de relações não-taxonômicas. Nesse sentido, o trabalho de Sanchéz e Moreno [29] se assemelha a proposta dessa dissertação, pois em ambos os verbos extraídos são utilizados para identificar as relações não-taxonômicas entre os conceitos.

A primeira tarefa inicia-se ainda na etapa de aprendizagem de taxonomia, quando são extraídos os verbos encontrados na mesma sentença que a palavra chave do domínio. Porém, devido a variação de forma dos verbos e a escolha de uma abordagem automática e sem supervisão, o sistema não consegue detectar a forma semântica de sujeito, verbo e objeto em uma sentença. Por isto, são extraídos somente verbos na forma assertiva. Nesse sentido, o processo proposto nessa dissertação tem vantagens frente ao trabalho de Sanchéz e Moreno [29], pois parte-se de um conjunto de termos que possui diversas informações associadas, como etiquetas sintáticas, semânticas, função gramatical, etc.

O processo de Sanchéz e Moreno [29] continua atuando sobre os verbos selecionados são então classificados em sucessor ou antecessor, de acordo com a posição em que aparecem relacionados a palavra-chave do domínio na sentença. A próxima etapa consiste em identificar os verbos que estão mais relacionados ao domínio de pesquisa. Para isto, é calculado um índice para avaliar a relevância de um verbo relacionado a palavra-chave do domínio. Este índice é calculado de forma simples através da razão entre a frequência do verbo junto da palavra-chave do domínio e da frequência apenas do verbo.

Os verbos são então classificados de acordo com o índice calculado, que permite avaliar quais são os verbos mais relacionados ao domínio da pesquisa. Em um estudo de caso, utilizando a palavra-chave hipertensão, observou-se a frequência do verbo com a palavra-chave e da frequência apenas da palavra-chave definindo conjuntos com o objetivo de criar um ponto de corte. Através deste ponto de corte foram eliminados cerca de 70% dos verbos extraídos.

Após selecionar os verbos mais frequentes, são feitas queries de busca com o objetivo de encontrar conceitos relacionados a palavra-chave através destes verbos. Somente são considerados candidatos a relação não-taxonômica do domínio os resultados das buscas que compreendem a estrutura <sujeito> <verbo> <objeto>, podendo a palavra-chave assumir a função de sujeito ou de objeto de acordo com a classificação do verbo (sucessor ou antecessor).

Com o objetivo de encontrar as relações não-taxonômicas mais relacionadas ao domínio de pesquisa, é calculado um índice através da razão entre a frequência da palavra-chave e do conceito encontrado (não necessariamente na mesma sentença) e da frequência do conceito. Para Sanchéz e Moreno [29], o ponto de corte nessa etapa deve ser menor que o da etapa anterior.

De acordo com Sanchéz e Moreno [29], não há um padrão considerado ótimo para avaliar relações não-taxonômicas e devido a grande quantidade de relações extraídas também torna-se difícil avaliá-las de forma manual. Com isto, a avaliação é realizada através de medidas de similaridade entre conceitos obtidas através da WordNet. Embora a WordNet não apresente resultados da relação entre conceitos através de um verbo, utiliza-se os valores atribuídos a similaridade entre dois conceitos.

Através dos índices obtidos, verifica-se através da análise de classificação em falsos positivos, ou no caso relações não selecionadas que são consideradas corretas ou incorretas. Este método de avaliação é influenciado pela disponibilidade dos conceitos encontrados na WordNet.

Se sugere a utilização de índices de avaliação de precisão e de recall. Nesta avaliação, relações não-taxonômicas do domínio câncer obtiveram resultados significantemente melhores que os obtidos no domínio hipertensão. Isto se deve pela maior disponibilidade de material do domínio câncer na Web [29].

De acordo com Sanchéz e Moreno [29], a proposta diferencia-se das demais pelo fato de iniciar o processo através da busca dos verbos relacionados ao domínio que serão candidatos a identificados da relação não-taxonômica. É válido ressaltar, que devido as dificuldades para avaliar corretamente as relações não-taxonômicas extraídas, é possível a inclusão da avaliação de um especialista do domínio ao final do processo.

Assim como nessa dissertação e no trabalho de Sanchéz e Moreno [29] a proposta apresentada por Villaverde *et al*. [33] baseia-se na premissa de que as relações não-taxonômicas são geralmente expressas e identificadas por verbos que relacionam pares de conceitos. A proposta tem como fonte um *corpus* de domínio, uma lista de candidatos a conceitos ou uma hierarquia de conceitos que descreva as relações taxonômicas presentes nos textos do *corpus*. O objetivo da

proposta [33] é extrair relações entre os conceitos e recomendá-las a engenheiros de ontologias e especialistas de domínio para que estes possam validá-las e adicioná-las a ontologia (Figura 4).

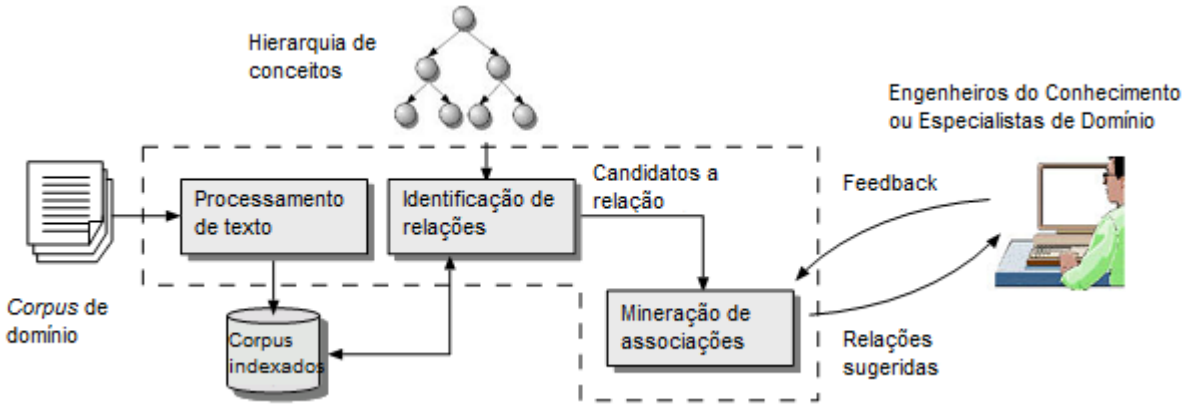

Figura 4. Processo de extração de relações não-taxonômicas de Villaverde *et al.* [33]

Fonte: Adaptado de Villaverde *et al.* [33]

Na primeira etapa do processo, assim como nessa dissertação, através de pré-processamento é feita a eliminação de *stop-words,* que são palavras que não acrescentam informações relevantes ao processo. Após a eliminação das *stop-words* os textos não-estruturados são transformados em representações adequadas para serem utilizadas como fonte para algoritmos de aprendizagem de máquina.

Os termos resultantes do processamento então indexados juntamente com o valor do índice TF-IDF [28] que representa sua relevância no texto em que foi extraído. Ainda na primeira etapa, é criado um conjunto de sinônimos de conceitos de uma ontologia através da WordNet. De acordo com Villaverde *et al*. [33] a utilização de sinônimos auxilia muito na extração de conceitos de um *corpus* de domínio.

Na segunda etapa do processo, é realizada uma busca no *corpus* de domínio com o objetivo de identificar pares de conceitos relacionados por um verbo em uma mesma sentença. Os resultados desta busca formam uma tripla definida por <conceito 1, conceito 2, verbo > e são agrupados em um conjunto de candidados a relacionamentos que são posteriomente recomendados para avaliação do especialista de domínio.

Sobre este conjunto é aplicado um algoritmo de mineração de regras de associação definidas por { < c*i* , c*j* > → *v* | c*i* , c*j* Є C e *v* é um verbo }, sendo C o conjunto de conceitos de um domínio. Como resultado são extraídas regras que são então avaliadas de acordo com evidências estatísticas, como suporte e confiança. Caso o suporte à relação seja considerado suficiente recomenda-se a relação para avaliação dos engenheiros de ontologias e especialistas de domínio.

A avaliação das relações não-taxonômicas extraídas no trabalho de Villaverde *et al.* [33] é realizada através da comparação entre as relações extraídas pelo processo automático e as relações não-taxonômicas definidas em uma ontologia construída de forma manual. Conduziu-se um experimento que utilizou o *corpus* Genia e sua respectiva ontologia. O *corpus* contém cerca de 1000 abstracts de artigos da base de dados *Medline* e entorno de 400.000 palavras. Por sua vez, a ontologia construída com base neste *corpus* apresenta 47 categorias relevantes no domínio da biologia.

Na realização do experimento, foram encontrados 304 padrões, cada um consistindo em um par de conceitos e o verbo que os relaciona. Deste total de relações encontradas 77% expressam relações válidas no domínio da biologia, enquanto 33% foram consideradas inválidas. Por relações válidas, considerou-se relações com índices de confiança aceitáveis para recomendação a especialistas de domínio.

De acordo com Villaverde et al. [33], o processo proposto auxilia na diminuição de sobrecarga sobre os especialistas de domínio e engenheiros de conhecimento no processo construção de ontologias. Embora ainda sejam necessários mais experimentos, considerou-se um bom resultado o número de relações não-taxonômicas recomendadas para avaliação dos especialistas.

O processo proposto por Maedech e Staab [21] diferencia-se do presente trabalho, mas também dos trabalhos de Sanchéz e Moreno [29] e Villaverde *et al.* [33]. A diferença está em extrair relações não-taxonômicas sem utilizar o verbo como elemento principal e por ter seu foco na língua alemã. O processo utiliza *corpus* de domínio como fonte, e está incluído no processo de aprendizagem de taxonomia da ontologia.

Maedech e Staab [21] afirmam que ao se focar no verbo muitas relações não-taxonômicas presentes em um *corpus* são negligenciadas. Porém, nessa dissertação o uso de conceitos de domínio e suas informações contextuais permite a extração de relações não-taxonômicas entre conceitos conhecidamente relevantes para o domínio e facilita a descoberta de relações entre eles.

O processo proposto por Maedech e Staab [21], utiliza métodos mais superficiais (especificamente *shallow processing*) em textos para extrair padrões linguísticos de relações entre conceitos. Um algoritmo de descoberta de regras de associação generalizadas analisa a saída do processo anteriormente executado e utiliza o conhecimento prévio obtido pela taxonomia da ontologia para sugerir relações entre conceitos pertencentes as suas classes.

Tendo como exemplo a frase : "O custo do albergue da juventude aumentou para R$20,00 por noite", o processo linguístico descobre que a palavra "custo" ocorre com frequência com palavras como "albergue", "casa de hóspedes" e "albergue da juventude". O algoritmo de descoberta determina medidas de suporte e confiança para os pares formados por estas palavras, bem como para palavras similares a estas identificadas pelas classes na ontologia, como "acomodação" e "custo". Na última etapa, o algoritmo determina em qual nível da ontologia deve ser adicionada a relação descoberta pela sentença dada.

Com o objetivo de avaliar o processo proposto, foi realizado um experimento com um *corpus* de 2234 documentos HTML, com aproximadamente 16.000 palavras de domínio do turismo. Através do processo proposto, foram extraídas 51.000 pares de palavras relacionadas não-taxonomicamente. Estes pares foram comparados a uma ontologia modelada previamente sobre o mesmo domínio que continha 284 conceitos e 88 relações não-taxonômicas.

Os dois conjuntos de relações não-taxonômicas foram analisados estatisticamente através do cálculo de precisão, recall e um índice específico desenvolvido para estimar a diferença entre as relações não-taxonômicas construídas automaticamente e manualmente. Na execução do experimento, o melhor valor de *recall* e precisão foi de 13% e 11% respectivamente. Com isso verificou-se que o processo proposto por Maedech e Staab [21] é fraco quanto a extração automática de relações não-taxonômicas, porém é um recurso válido para auxiliar na construção manual de relações não-taxonômicas. Neste caso, pode-se verificar que as abordagens que tem o

verbo como elemento central para identificação de relações não-taxonômicas, incluindo este trabalho, apresentam resultados mais significativos.

O trabalho de Serra e Girardi [31] apresenta um processo para extração semi-automática de relações não-taxonômicas entre conceitos de duas classes em textos da língua inglesa (Figura 5). Assim como nessa dissertação, nos trabalhos de Sanchéz e Moreno [29] e Villaverde *et al.* [33], o método extrai as relações indicadas por verbos nas sentenças e sugere a possível melhor classificação hierarquica em que a relação pode ser adicionada. O processo é dividido em três etapas: (i) extração de candidatos a relação não-taxonômica; (ii) análise do nível hierárquico; e (iii) seleção manual das relações.

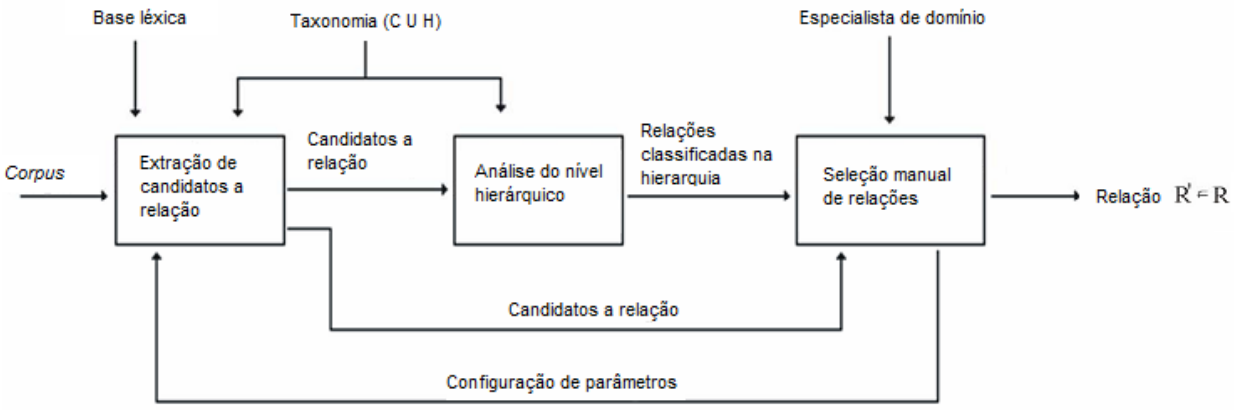

Figura 5. Proposta de extração de relações não-taxonômicas de Serra e Girardi [31]

Fonte: Adaptado de Serra e Girardi [31]

A etapa de extração de candidatos a relação não-taxonômica utiliza técnicas de PLN para extrair um conjunto inicial de relações a partir de um *corpus* de domínio. Inicialmente o texto é dividido em sentenças, pois serão identificadas apenas relações explícitas, assim como se faz nessa dissertação e nos trabalhos de Sanchéz e Moreno [29] e Villaverde *et al.* [33].

Sobre o resultado são realizadas buscas com o objetivo de encontrar sentenças que contenham dois termos que representam conceitos da hierarquia de classes de uma ontologia. As classes identificadas são então expandidas, como objetivo de realizar buscas de relações não-taxonômicas também a respeito dos sinônimos dos conceitos encontrados. O último passo desta

etapa se foca em encontrar verbos que conectem os conceitos nas sentenças extraídas na fase anterior, e os agrupa em uma tripla definida por < *conceito 1, verbo, conceito 2* >.

O processo também extrai relações não-taxonômicas classificadas como independentes de domínio, identificadas na língua inglesa geralmente pelo apóstrofo. Quando ocorre essa situação, a tripla é formada pelos dois conceitos e o verbo identificado pelo apóstrofo é substituído pelo verbo "has". Neste caso será posteriormente sugerido ao especialista de domínio que defina um novo verbo que representa a relação entre os conceitos. Esta situação está fora do escopo dessa dissertação, pois a classificação de relações em dependentes ou independentes de domínio aplica-se de forma mais esclarecida na língua inglesa.

Na segunda etapa, que é opcional, é utilizado um algoritmo para descoberta de regras de associação generalizadas com o objetivo de classificar as relações não-taxonômicas da melhor forma na hierarquia da ontologia. Na última etapa especialistas de domínio são responsáveis por avaliar os resultados encontrados antes de adicioná-los a uma ontologia, pois de acordo com Serra e Girardi [31] nenhuma técnica de PLN ou Aprendizagem de Máquina substitui a decisão de um especialista de domínio quanto a avaliação da relevância de uma relação não-taxonômica.

Assim como messa dissertação, com o objetivo de automatizar o processo de extração de relações não-taxonômicas extraídas, no trabalho de Serra e Girardi [31] uma ferramenta computacional foi desenvolvida. Para que seja possível avaliar o produto do processo, foi realizado um experimento do processo com um *corpus* de 500 textos do domínio de Direito da Família.

O resultado deste experimento será comparado a uma ontologia do mesmo domínio previamente construída. A efetividade do processo foi avaliada através de índices de precisão, *recall* e *f-measure*.

A proposta de Schutz e Buitelaar [30] se foca na língua alemã e inglesa, e assim como nessa dissertação e nos trabalhos de Sanchéz e Moreno [29], Serra e Girardi [31] e Villaverde *et al.* [33], utiliza o verbo como identificador de relações não-taxonômicas. O trabalho tem como fonte de dados relatórios minuto-a-minuto de partidas de futebol do campeonato alemão. O projeto foi desenvolvido no contexto da SmartWeb, que tinha por objetivo o desenvolvimento de aplicações para dispositivos móveis referentes a Copa do Mundo sediada na Alemanha no ano de 2006.

O trabalho apresenta um aplicativo, denominado RelExt, que tem por objetivo extrair automaticamente relações não-taxonômicas que podem ser adicionadas a uma ontologia já existente. Para isto, ele tem como fonte de dados um *corpus* de domínio e utiliza métodos linguísticos e estatísticos para extrair relações entre conceitos e verbos.

Através de métodos linguísticos e de uma ontologia de domínio de esportes é gerado um *corpus* anotado, no qual identifica-se a função sintática e a classe da ontologia a qual cada elemento da sentença pertence. Ainda neste processo são identificados palavras sinônimas que correspondem a mesma classe da ontologia e nomeadas instâncias conhecidas, como por exemplo identificar que "Oliver Kahn" pertence a classe "Goleiro".

Na etapa de processamento estatístico, verifica-se no *corpus* a frequencia de cada uma das classes no papel de sujeito, objeto direto e objeto indireto para cada verbo. Após identificar a frequencia das classes, é feita a composição das triplas < "Classe ', "verbo", "Classe 2> entre as classes que aparecem com mais frequencia. Por exemplo, verifica-se que a classe "Jogador de Futebol" aparece com mais frequência como sujeito do que "Juiz de Futebol" junto ao verbo "marcar", e a classe "Gol" aparece mais vezes que a classe "Time" na mesma situação. Neste caso, a relação não-taxonômica é identificada pela classe <"Jogador de Futebol", "marcar", "Gol">.

Com o objetivo de avaliar o processo proposto, os autores realizaram um experimento no qual foram organizados 4 conjuntos com 300 documentos de textos cada. Em cada um dos conjuntos foi executado o processo, resultando em um total de 192 triplas de relações não-taxonômicas. Estas triplas foram submetidas a avaliação de 3 especialistas de domínio, que classificaram 65% delas como relevantes para o domínio.

O trabalho de Weichselbraun *et al*. [35] se foca na adição de descoberta de relações entre conceitos a uma arquitetura de extenção semi-automática de ontologias. Esta arquitetura constrói ontologias específicas de domínio com base em uma pequena ontologia inicial e um *corpus* de domínio contendo uma grande quantidade de documentos web não estruturados. Assim como o processo apresentado nessa dissertação e nos trabalhos de Sanchéz e Moreno [29], Villaverde *et al.* [33], Serra e Girardi [31] e Schutz e Buitelaar [30], o processo proposto por Weichselbraun *et al*. [35] utiliza verbos como elementos centrais de conexão entre conceitos.

Em Weichselbraun *et al*. [35] distingue a descoberta de relações em taxonômicas e não taxonômicas. As relações taxonômicas são identificadas através da análise dos substantivos centrais de uma oração, identificação de sinônimos através da *WordNet* e análise de subordinação dos termos. Com relação a descoberta de relações não-taxonômicas é descrita uma abordagem *bottom-up* para sugestão automática de relações.

O processo proposto extrai um vetor de verbos de relações semânticas identificadas em um *corpus* de domínio, os agrupa identificando centróides de relações conhecidas e armazena os centróides em uma base de conhecimento. Compara-se então um vetor de verbos extraídos de relações desconhecidas com os centróides armazenados na base de conhecimento.

A similaridade entre os verbos de relações desconhecidas com os verbos registrados na base de conhecimento são computadas e os que apresentarem maior similaridade são sugeridos para o especialista de domínio como relação não-taxonômica (Figura 6). As relações classificadas como relevantes são então adicionadas a ontologia previamente criada.

Com o objetivo de avaliar o processo proposto, utilizou-se um *corpus* com cerca de 200.000 documentos extraídos de diretórios de sites de notícias locais. Ao final do processo de construção de ontologia, restaram 17 relações entre conceitos que não haviam sido identificadas e estas foram então utilizadas para o experimento com o processo de descoberta de relações não-taxonômicas.

Neste processo, utilizaram-se duas fontes para descoberta das relações não-taxonômicas que deveriam ser sugeridas aos especialistas, uma online e outra local. De acordo com Weichselbraun *et al*. [35], o método que tinha como fonte a Web apresentou melhores resultados devido a vasta quantidade de conteúdo que se encontra na internet.

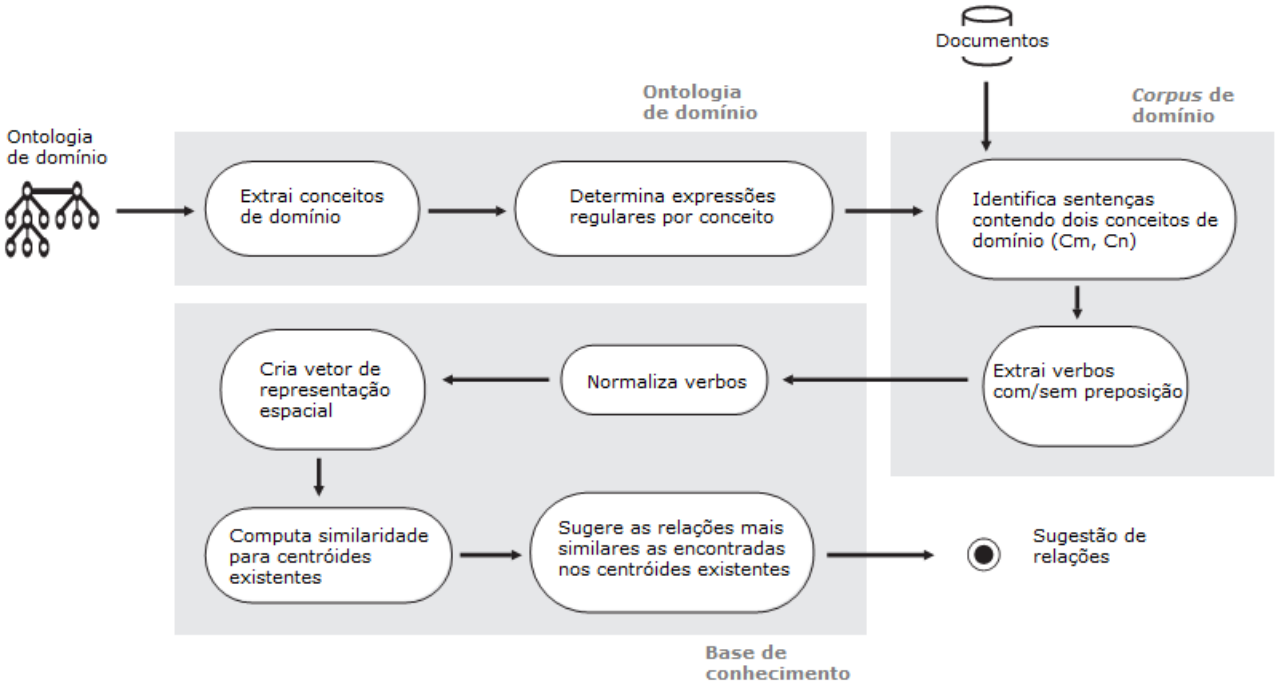

Figura 6. Arquitetura para sugestão automática de relações não-taxonômicas

Fonte: Adaptado de Weichselbraun *et al*. [35]

A avaliação se deu através da quantidade de sugestões que o sistema fez até que a relação correta fosse identificada. Na primeira tentativa de sugestão o melhor resultado foi através da base local com 55% de acerto e com mais de uma tentativa o melhor resultado foi através da base online com 79,4% de acerto. A Tabela 3 apresenta uma comparação dos trabalhos similares ao proposto nessa dissertação.

Tabela 3. Comparação de trabalhos similares

Autor	**Fonte de dados**	**Relação identificada por verbo?**	**Idioma**	**Tipos de relações não-taxonômicas extraídas**
Sanchéz e Moreno [29]	Web	Sim	Inglês	Dependente de domínio
Villaverde *et al.* [33]	*Corpus* de domínio, lista de candidatos a conceitos ou hierarquia de conceitos	Sim	Inglês	Dependente de domínio
Maedech e Staab [21]	*Corpus* de domínio	Não	Alemão	Dependente de domínio
Serra e Girardi [31]	*Corpus* de domínio	Sim	Inglês	Dependente e independente de domínio
Schutz e Buitelaar [30]	*Corpus* de domínio	Sim	Inglês e Alemão	Dependente de domínio
Weichselbraun *et al.* [35]	Ontologia e *corpus* de domínio	Sim	Inglês	Independente de domínio

Além dos trabalhos já citados podem ser mencionados ainda os processos propostos por trabalhos de Filkenstain e Morin [12], Byrd e Ravin [6], Kavalec e Svátek [17], e Nabila *et al.* [24].

O processo proposto por Filkenstain e Morin [12] combina métodos com supervisão e sem supervisão para extração de relações não-taxonômicas. O método sem supervisão utiliza técnicas de PLN para extrair adjetivos ou verbos que conectem conceitos encontrados em um *corpus*, enquanto o método supervisionado propõe relações padrão para determinadas entidades da ontologia.

No processo proposto por Byrd e Ravin [6], as relações entre os conceitos são identificadas através da utilização de automatos finitos construídos sobre os padrões das sentenças. Cada relação recebe um valor calculado através da frequência com que ocorre na tripla formada pelos dois conceitos e o elemento que os conecta. Nesta proposta, o elemento central não é necessariamente um verbo.

De forma similar ao processo apresentado nessa dissertação e nos trabalhos de Sanchéz e Moreno [29], Villaverde *et al.* [33], Serra e Girardi [31], Weichselbraun *et al.* [35] e Schutz e Buitelaar [30], o processo proposto por Kavalec e Svátek [17] utiliza verbos como elemento central de conexão entre conceitos na identificação de uma relação não-taxonômica. No trabalho de Kavalec e Svátek [17] as relações são classificadas pela frequência com que aparecem em um *corpus* de domínio.

No trabalho de Nabila *et al.* [24], os verbos também assumem papel fundamental na identificação das relações não-taxonômicas em um *corpus*. Porém, o foco está em descobrir relações não-taxonômicas em sentenças diferentes. Neste processo, são extraídos verbos e conceitos de um *corpus* através de processos linguísticos. Os verbos com significado similar são agrupados, e a cada grupo também são adicionados os conceitos que se relacionam através de cada verbo. Após isso, é feito o produto cartesiano entre os sujeitos e objetos mais frequentes de um grupo de verbos. De acordo com Nabila *et al*. [24], esta proposta permite extrair relações não-taxonômicas em sentenças diferentes. Embora apresente maior quantidade de relações extraídas, a identificação das relações ainda precisa ser aprimorada.

3 DESENVOLVIMENTO

O Capítulo 3 apresenta o processo para extração de relações não-taxonômicas proposto nesse trabalho e detalha cada uma das etapas que o compõe. Ainda nesste Capítulo é apresentada a ferramenta desenvolvida para a visualização das relações não-taxonômicas extraídas.

3.1 Arquitetura da Solução

Esta seção apresenta uma visão geral do processo proposto neste trabalho para extração de relações não-taxonômicas. Este processo tem como fonte de dados a tabela de termos e conceitos extraídos a partir de *corpus* em língua portuguesa obtida através do ExATOlp [20]. Essa tabela também contém informações contextuais para cada um dos conceitos (Tabela 1, Capítulo 2, Seção 2.2, pg. 22).

Embora já existam propostas para extração de relações não-taxonômicas a partir de *corpus* de domínio, nenhuma se aplica a língua portuguesa. Além disso, nenhuma proposta utiliza como fonte de dados os conceitos em conjunto com suas informações contextuais.

Diferente das propostas existentes, o processo proposto nesse trabalho tem como ponto de partida conceitos já definidos de um domínio. Com isso, uma etapa para extração de conceitos diretamente de um *corpus* não é necessária, permitindo então que o processo tenha seu foco somente na extração de relações não-taxonômicas.

A Figura 7 apresenta uma esquematização sobre o processo de extração de relações não-taxonômicas. O processo é composto de 5 etapas: (i) Aquisição de termos de domínio; (ii) Eliminação de termos com informações faltantes; (iii) Identificação de conceitos; (iv) Extração de relações não-taxonômicas; e (v) Visualização de relações não-taxonômicas. Estas etapas são detalhadas a seguir.

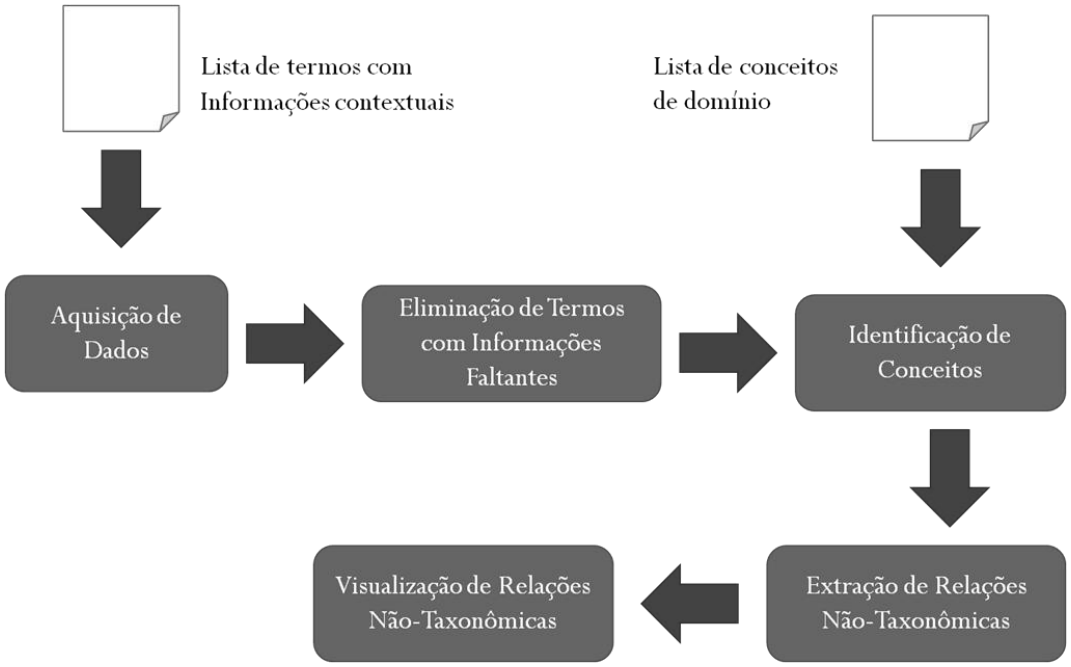

Figura 7. Visão geral do processo proposto

3.1.1 Aquisição de Termos de Domínio

A etapa de aquisição consiste em adquirir os termos de domínio que serão utilizados como fonte para o processo. Conforme pode ser visualizado na Figura 7, o processo tem como fonte de dados os termos de domínio e suas informações contextuais obtidos através do software ExATOlp [20].

Diferente dos processos propostos nos trabalhos similares, o processo proposto nesse trabalho não se responsabiliza por extrair termos e classificá-los em conceitos de um domínio. Como os termos são obtidos através do processo de extração de conceitos proposto por Lopes [19], a etapa de aquisição está relacionada apenas em obter os dados da saída desse processo.

3.1.2 Eliminação de Termos com Informações Faltantes

A segunda etapa do processo consiste na eliminação de todos os termos para os quais não foram extraídas ou identificadas as informações contextuais pertinentes ao processo, sendo essas identificadas pelo número correspondente na tabela gerada pelo ExATOlp (Tabela 1, Capítulo 2, Seção 2.2, pg. 22): (i) conceito na forma canônica (2); (ii) função gramatical do conceito (4); (iii) identificador da frase de onde o conceito foi extraído (13); (iv) identificador do documento de onde o conceito foi extraído (14); e o predicado ao qual o conceito exerce sua função gramatical (17). Apesar de eliminar termos, essa etapa é diferente dos trabalhos que fazem elininação de stopwords, como é o caso dos processos propostos por Villaverde et al. [33], Sanchéz e Moreno [29] e Nabila *et al.* [24].

Nesses trabalhos similares, como há também o foco na extração de conceitos de domínio, a etapa de eliminação de stopwords é considerada uma etapa fundamental. Através dela são eliminados termos que, ainda que frequentes, não correspondam a conceitos de um domínio. Dentre os termos eliminados pode-se citar palavras muito frequentes que não acrescentam informações sobre o domínio desejado [24, 33], advérbios, preposições e verbos que indicam relações taxonômicas (por exemplo, os verbos “é”, “são”, “inclui”) [29].

No processo proposto nessa dissertação, além dos termos com informações contextuais faltantes, são eliminados também todos os termos que tem função sintática diferente de sujeito e objeto, identificada no item 4 da Tabela 1 (Capítulo 2, Seção 2.2, pg. 22). Utiliza-se somente as informações contextuais dos termos e conceitos de domínio, nesse sentido, a etapa de eliminação é necessária. Pois torna-se impossível realizar o processo utilizando os termos que não possuem as informações contextuais necessárias.

3.1.3 Identificação de Conceitos

A terceira etapa do processo consiste na eliminação de todos os termos que não são conceitos de domínio. Para isso, comparam-se todos os termos que restaram após a segunda etapa do processo com a lista de conceitos de domínio produzida pela ferramenta ExATOlp [20]. Todos os termos que não são considerados conceitos para o domínio em questão são, desta forma, eliminados.

Cada um dos conceitos extraídos pelo ExATOlp [20], possui um valor numérico diretamente relacionado a sua relevância no domínio do qual foi extraído. Este valor numérico é obtido através do índice *tf-dcf* [19] (Equação 1), que possibilita a identificação dos termos mais relevantes de um *corpus*, tornando-os então, conceitos do domínio.

De acordo com Lopes [19], a frequência absoluta de um termo em um *corpus* de domínio é a base do índice *tf-dcf* para considerar a sua relevância em um domínio. Em seguida, os termos que aparecem em *corporas* contrastantes são penalizados através da razão entre a frequência absoluta do termo no *corpus* de domínio e a composição geométrica da sua frequência absoluta em cada um dos *corpora* constrastantes. Com isso, a ocorrência em outros *corpora*, dimunui a relevância do termo no *corpus* de domínio que está sendo analisado.

Equação 1 – Índice *tf-dcf*
Fonte: Lopes [19]

$$tf\text{-}dcf_t^{(c)} = \frac{tf_t^{(c)}}{\prod_{\forall\, g \in \mathcal{G}} 1 + \log\left(1 + tf_t^{(g)}\right)}$$

Nos trabalhos similares essa etapa de identificação de conceitos está presente no início do processo, pois trata-se da atividade de extração de conceitos em um *corpus* de domínio. Embora o processo proposto nesse trabalho utilize termos e conceitos produzidos por outro processo [19], essa etapa é essencial. Pois o objetivo do trabalho é identificar relações não-taxonômicas somente entre conceitos de um domínio.

3.1.4 Extração de Relações Não-Taxonômicas

A quarta etapa do processo consiste na extração de relações não-taxonômicas entre os conceitos de um domínio. A realização dessa etapa segue duas abordagens: (i) relações não-taxonômicas geradas por produto cartesiano; (ii) extração de relações não-taxonômicas expícitas.

A **primeira abordagem**, assim como no trabalho de Nabila *et al.* [24], extrai relações não-taxonômicas a partir do produto cartesiano entre os conceitos que se relacionam através de um

mesmo verbo na forma canônica, identifcado pelo item 17 da Tabela 1 (Capítulo 2, Seção 2.2, pg. 17). Através desta abordagem é possível descobrir relações não-taxonômicas que ocorrem entre conceitos relacionados por verbos similares, que na maiora das vezes não aparecem em uma mesma sentença [24].

Nesta abordagem o primeiro passo é identificar todos os conceitos que se relacionam a um mesmo verbo. Este conjunto de conceitos é então classificado de acordo com a sua função sintática, sujeito ou objeto. As triplas candidatas a relação não-taxonômica do domínio são então geradas com o produto cartesiano dos conceitos sujeitos e conceitos objetos relacionados por um mesmo verbo.

A **segunda abordagem**, assim como nos trabalhos de Sanchéz e Moreno [24], Villaverde *et al* [33], Serra e Girardi [31], Schutz e Buitelaar [30] e Weichselbraun *et al.* [35], extrai relações não-taxonômicas entre conceitos que ocorrem em uma mesma sentença (explícita), que são sujeito e objeto do mesmo verbo. Através desta abordagem, não se pode garantir que todas relações são relevantes para o domínio, porém é possível afirmar que a chance delas serem relevantes é maior do que na primeira abordagem. Nessa abordagem, após a identificação dos conceitos do domínio, são identificados todos os conceitos que se relacionam através de um mesmo verbo na forma canônica, identificado pelo item 17 da Tabela 1 (Capítulo 2, Seção 2.2, pg. 22). São então consideradas triplas de relações não-taxonômicas do domínio, as relações que possuem conceitos pertencentes a mesma sentença. Assim como trabalhos similares, as triplas são definidas: <*conceito sujeito, verbo, conceito objeto*>.

Nos processos de extração de relações não-taxonômicas propostos por Sanchéz e Moreno [29], Serra e Girardi [31] e Villaverde *et al.* [33], para cada uma das relações extraídas é atribuído um valor numérico que busca indicar a relevância da relação para o domíndio em questão. Através do valor do índice obtido (precisão, confiança) são classificadas, ordenadas e selecionadas as relações que serão utilizadas, ou sugeridas, para avaliação dos especialistas do domínio.

Para que seja possível estimar a relevância das relações não-taxonômicas extraídas através do processo proposto, são obtidos dois valores numéricos através dos índices: (i) de frequência acumulada; e (ii) de frequência compartilhada. O índice *tf-dcf* de cada conceito é utilizado como base para obtenção dos referidos índices.

O índice de frequência acumulada é obtido através do somatório do índice *tf-dcf* de todos os conceitos relacionados através de uma mesma relação (Equação 2). Este índice tem por objetivo estimar a relevância de uma relação com base na relevância dos conceitos que ela relaciona, sejam eles sujeitos ou objetos. Com isto, uma relação é considerada mais relevante para o domínio conforme maior for a relevância dos conceitos que ela relaciona.

Equação 2. Índice de frequência acumulada

$$facum_r = \sum_{\forall r | \exists (t_1, r, t_2)} tf\text{-}dcf_{t_1}^{(c)} + tf\text{-}dcf_{t_2}^{(c)}$$

A Equação 2 apresenta a definição do índice de frequência acumulada, na qual (*t1*, *r*, *t2*) representa uma ocorrência da relação indicada por um verbo *r*, *t1* e *t2* representam, respectivamente os conceitos que são sujeito e objeto dessa ocorrência da relação. É importante ressaltar que o valor do índice *tf-dcf* de todos os conceitos envolvidos deve ser considerado, mesmo se o conceito se repetir ou ocorrer exercendo função de sujeito e objeto.

Na Tabela 4 é apresentado como exemplo o cálculo do índice de frequência acumulada para a relação "receber". Essa relação conecta o conceitos "lactente" com "estimulação" e "leucemia" com "quimioterapia". Através da soma do valor do índice *tf-dcf* para esses conceitos, é obtido o valor 167,50 para o índice de frequência acumulada para a relação "receber".

Tabela 4. Exemplo de cálculo de índice de frequencia acumulada

Sujeito (tf-dcf)	**Relação**	**Objeto (tf-dcf)**
Lactente (114,00)	receber 167,50	Estimulação (11,00)
Leucemia (14,50)	receber 167,50	Quimioterapia (28,00)

O índice de frequência compartilhada é obtido através da razão entre o índice de frequência acumulada de uma relação e a quantidade de vezes que a relação ocorreu (Equação 3). Através deste índice é possível estimar a relevância da relação não apenas pela relevância dos conceitos que ela relaciona, mas também levando em conta o número de vezes que ela ocorre.

Equação 3. Índice de frequência compartilhada

$$fcompart = \frac{facum_r}{\sum_{\forall r | \exists (t_1, r, t_2)} 1}$$

A equação 3 apresenta a normalização do índice de frequência compartilhada pelo número total de ocorrências da relação representada por *r*. Para obter o número de vezes que a relação ocorre deve-se considerar a quantidade de vezes que um mesmo verbo aparece, sem levar em conta os conceitos relacionados a ele. Com isto, uma relação mais relevante é aquela que relaciona conceitos mais relevantes, sem necessariamente ocorrer muitas vezes.

Através do exemplo apresentado na Tabela 4, é possível obter o valor do índice de frequência compartilhada para a relação "receber". Esse valor é obtido através da razão entre o valor do índice de frequencia acumulada (167,50) e o número de vezes em que a relação é encontrada, no caso do exemplo apresentado, duas vezes. Com isso, o valor do índice de frequência compartilhada para a relação receber é 83,75.

3.2 Visualização de Relações Não-Taxonômicas

A quinta e última etapa do processo tem como objetivo permitir a visualização das relações não-taxonômicas extraídas. Essa etapa é essencial para que as relações possam ser exploradas e manipuladas de forma simplificada.

Embora nenhum dos trabalhos similares apresente uma etapa do processo que especifique a visualização das relações extraídas, Villaverde *et al* [33] e Schutz e Buitelaar [30] apresentam-nas aos especialistas através de uma interface Web. Nas próximas seções sera apresentada a ferramenta Web desenvolvida para permitir a visualização e manipulação das relações não-taxonômicas extraídas através do processo proposto.

3.2.1 Ferramenta Proposta

Para visualizar as relações extraídas foi desenvolvida uma ferramenta Web na linguagem de programação PHP e que utiliza o SGBD (Sistema Gerenciador de Banco de Dados) MySQL. Através deste aplicativo é possível explorar e manipular as relações extraídas.

A Figura 8 apresenta a tela inicial do aplicativo. Nessa tela é apresentada uma visão geral do trabalho realizado, informações sobre os autores, publicações realizadas sobre o trabalho e informações sobre a utilização do aplicativo. Dentre as informações disponibilizadas sobre o aplicativo estão a descrição dos *corpora* de domínio utilizados, uma pequena amostra dos textos de cada um dos *corpus* e informações sobre o Índice de Frequência Acumulada, utilizado para classificar as relações no aplicativo.

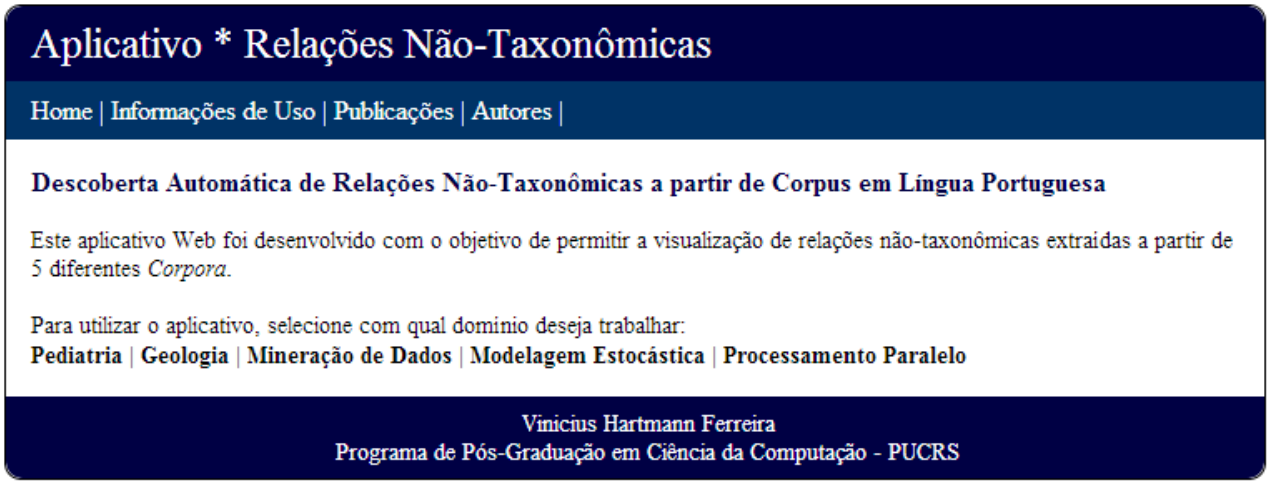

Figura 8. Tela inicial do aplicativo desenvolvido

Para explorar e manipular as relações não-taxonômicas o primeiro passo é selecionar o *corpus* de domínio. Após isso, é apresentada uma tela que permite exibir e organizar por ordem alfabética os sujeitos, relações ou objetos do domínio selecionado (Figura 9). É também possível ordenar os sujeitos e objetos mais frequentes de acordo com o seu valor de índice de frequência *tf-dcf*. Por sua vez as relações também podem ser ordenadas por frequência de acordo com seu valor de Indice de Frequência Acumulada, tendo o valor do Índice de Frequência Compartilhada como critério de desempate.

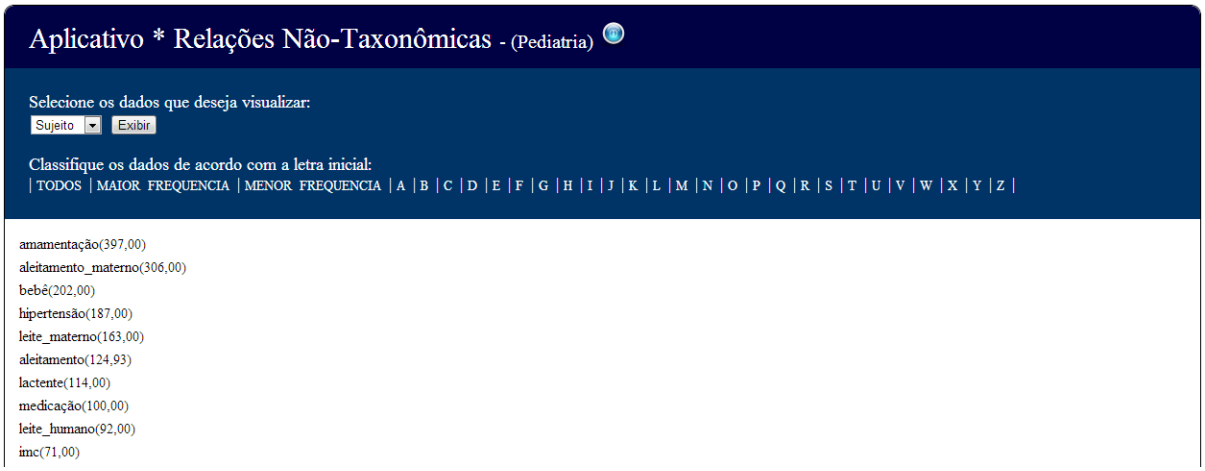

Figura 9. Aplicativo apresentando sujeitos do *corpus* de domínio de Pediatria

Ao selecionar um conceito, ou relação, é apresentada a interface que detalha as relações não-taxonômicas relacionadas ao item selecionado. Por exemplo, ao selecionar o conceito sujeito "ascaridíase", são detalhadas todos as relações não-taxonômicas referentes a ele, conforme pode ser visto na Figura 10.

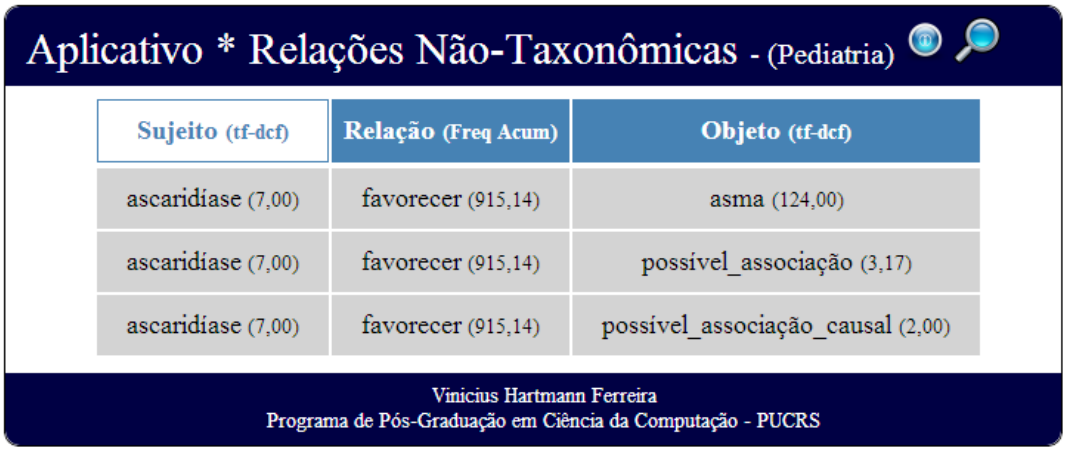

Figura 10. Relações não-taxonômicas referentes ao conceito ascaridíase

Ao lado dos conceitos são apresentados os seus valores de índice *tf-dcf*, e ao lado das relações (verbos) é apresentado o seus valores de índice de Frequência Acumulada. Ainda na tela apresentada na Figura 10, ao clicar em qualquer um dos conceitos ou verbos, é possível obter o detalhamento de suas relações não-taxonômicas. Por exemplo, na Figura 11 é apresentado o detalhamento das relações não-taxonômicas referentes a relação "favorecer".

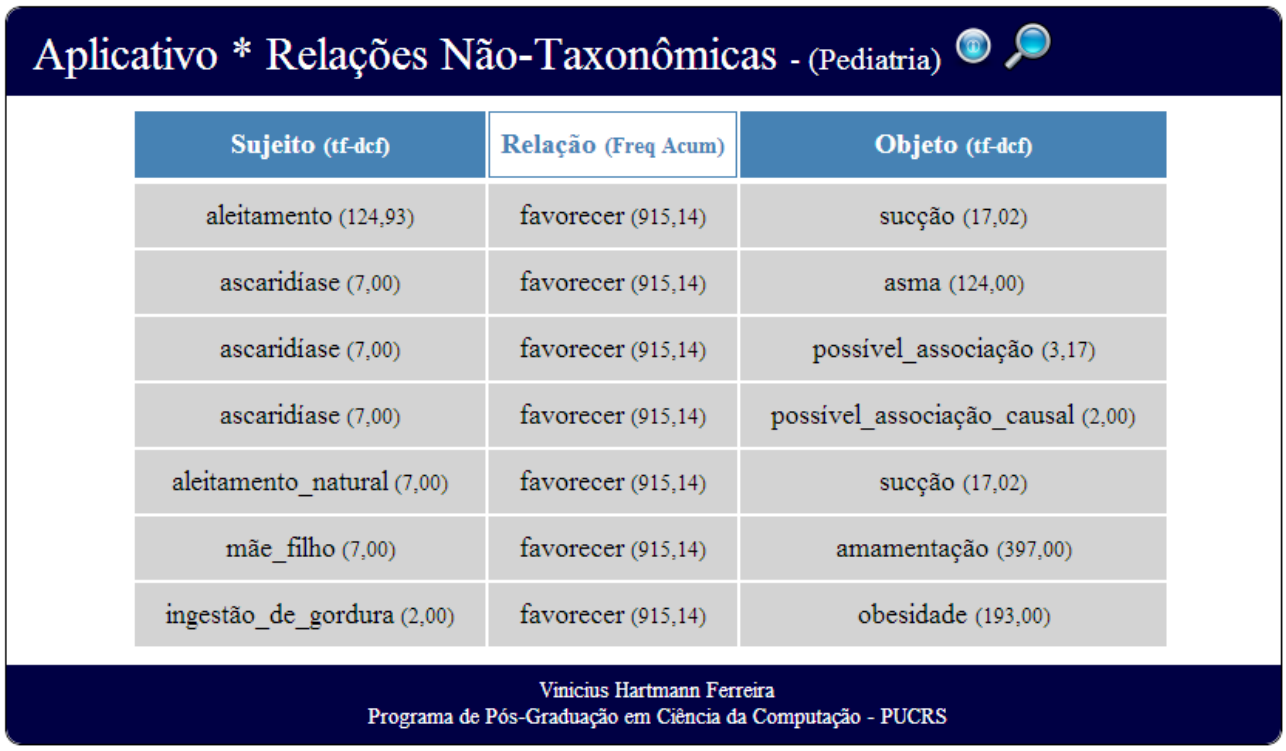

Figura 11. Detalhamento das relações não-taxonômicas para a relação "favorecer"

Na próxima seção serão apresentadas e detalhadas as opções do menu de navegação disponível na ferramenta.

3.2.2 Menus de Navegação

A ferramenta de visualiação de relações não-taxonômicas apresenta quatro opções no menu de navegação: (i) Tela inicial; (ii) Informações de uso; (iii) Autores; (iv) Publicações.

A tela inicial, conforme pode ser visto na Figura 8, descreve qual é o objetivo da ferramenta. Além disso, também apresenta os *corpora* de domínio para os quais é possível visualizar as relações não-taxonômicas extraídas. É através da seleção dos *corpora* de domínio que o usuário tem acesso as suas relações não-taxonômicas.

A opção de informações de uso descreve os principais elementos envolvidos no processo de extração de relações não-taxonômicas. Esses elementos são divididos em: (i) Descrição dos *corpora*; (ii) Índice de relevância; e (iii) Amostra dos *corpora*.

Através da descrição dos *corpora*, são apresentadas informações referentes aos *corporas* para os quais foram extraídas relações não-taxonômicas. Dentre essas informações são apresentados o número de palavras, frases e documentos de cada *corpora*, além do número de termos e conceitos identificados em cada um.

Na opção de índice de relevância é descrito o Índice de Frequência Acumulada (Equação 2), utilizado para ordenar as relações na ferramenta. Nessa opção também é apresentado um exemplo de aplicação do Índice de Frequência Acumulada e uma breve descrição do índice *tf-dcf* [19], utilizado para ordenação e classificação dos conceitos de domínio.

Na opção de amostras dos *corpora*, para cada *corpus* é disponibilizado um arquivo com amostras de textos. Através destas amostras é possível visualizar de onde os termos, conceitos e informações contextuais foram extraídos.

Dentre as opções do menu de navegação ainda é possível acessar informações sobre os autores e sobre as publicações aceitas referentes a esse trabalho. Na próxima seção serão apresentados o estudo de caso realizado com o processo proposto, a avaliação e discussão dos resultados obtidos.

4 APLICAÇÃO E ANÁLISE

O Capítulo 4 apresenta os resultados obtidos através da aplicação do processo proposto em cinco *corpora* de domínio diferentes. Além disso também apresenta os resultados obtidos de análises feitas por especialistas sobre três diferentes aplicações: a de especificidade das relações, a de relevância dos conceitos e a de anotação de papéis semânticos. Dessa forma, esse Capítulo está organizado em duas seções: Caso de Estudo (4.1) e Análise (4.2).

4.1 Caso de Estudo

Com o objetivo de verificar o funcionamento do processo proposto, foi desenvolvido um sistema computacional em Java para operacionalizar as suas quatro primeiras etapas. Para este experimento foram utilizados como fonte de dados termos e conceitos extraídos pelo ExATOlp [20] a partir de

cinco *corpora* de domínio: (i) Pediatria; (ii) Geologia; (iii) Mineração de dados; (iv) Modelagem estocástica; e (v) Processamento paralelo.

O *corpus* de pediatria foi desenvolvido por Coulthard [19] a partir de 283 textos do Jornal de Pediatria, um periódico bilíngüe da Sociedade Brasileira de Pedriatria. Este c*orpus* foi cedido pelo projeto TEXTPED e sofreu um processo de refinamento que o deixou com 281 textos. Os *corpora* de domínio de Geologia, Mineração de dados, Modelagem estocástica e Processamento paralelo são compostos por textos científicos (Teses, Dissertações e Artigos) e foram construídos por Lopes [19] para suprir as necessidades de desenvolvimento de sua Tese de Doutorado.

Tabela 5. Corpora de domínio utilizados no experimento

Domínio	**Textos**	**Frases**	**Palavras**	**Termos**	**Conceitos**
Pediatria	281	27.724	835.412	180.120	8.270
Geologia	234	69.461	2.010.527	436.401	25.173
Mineração de dados	53	42.932	1.127.816	244.439	12.816
Modelagem estocástica	88	44.222	1.173.401	252.178	12.582
Processamento paralelo	62	40.928	1.086.771	241.145	11.591

Tabela 5 apresenta a síntese do conteúdo de cada um dos corpora utilizados na aplicação do processo. Através desta tabela é possível identificar a quantidade de textos, frases, palavras, termos e conceitos de cada um dos *corpora*.

4.1.1 Produto Cartesiano

A primeira abordagem utilizada para a extração de relações não-taxonômicas foi a partir do produto cartesiano entre os conceitos relacionados através de um mesmo verbo. A aplicação utilizando essa abordagem foi realizada apenas com o *corpus* de Geologia.

Através dessa aplicação foram extraídas 486.604 triplas (instâncias) de 429 relações não-taxonômicas distintas (verbos). O grande número de triplas encontradas se dá pelo fato de o produto cartesiano gerar triplas com todos conceitos sujeito e conceitos objeto relacionados a um verbo.

Conforme se verá nas próximas seções, o número de triplas e relações geradas pelo produto cartesiano é muito superior ao obtido pela abordagem de identificar relações explícitas, fato também ocorrido no trabalho de Nabila *et al.* [24]. Embora não seja possível afirmar que todas as relações obtidas na segunda abordagem sejam relevantes, existe uma probabilidade muito maior de que essas relações sejam mais relevantes do que as relações obtidas pela primeira abordagem, conforme verificado também por Nabila *et al.* [24].

Um experimento foi realizado nessa dissertação aplicando essa abordagem. Os resultados obtidos foram publicados no Seminário de Pesquisa em Ontologia no Brasil (ONTOBRAS 2012) [11]. Embora os resultados do artigo sejam promissores, o foco desse trabalho está em extrair relações não-taxonômicas explícitas de domínio ao invés de criar novas relações que podem não ser específicas de domínio, que é o que ocorre na abordagem por produto cartesiano.

Para conferir o quão próximas eram as relações obtidas pelos duas abordagens, foi realizado uma intersecção entre as triplas obtidas pelas duas abordagens com o *corpus* de Geologia. Este processo identificou que não havia nem 1% (0.38%) de intersecção entre elas. Com isto, a abordagem de extração por produto cartesiano foi descartada e optou-se pela abordagem que extrai relações explicitas, ou seja, identificadas por conceitos relacionados por um mesmo verbo em uma mesma sentença.

4.1.2 Relações Não-Taxonômicas Explícitas

A aplicação sobre os cinco *corpora* de domínio foi realizada utilizando a abordagem de extração de relações não-taxonômicas explícitas. Utilizando o sistema computacional desenvolvido, foram executadas as quatro primeiras etapas do processo proposto. Na etapa de aquisição, o sistema computacional obteve os termos e informações contextuais de cada um dos *corpus* de domínio (Tabela 6).

Tabela 6. Termos obtidos na primeira etapa do processo

Domínio	Termos
Pediatria	180.120
Geologia	436.401
Mineração de dados	244.439
Modelagem estocástica	252.178
Processamento paralelo	241.145

Após obter os termos produzidos pelo ExATOlp [20], a etapa seguinte do processo consiste da eliminação dos termos que possuem informações contextuais necessárias faltantes ou que não possuam função gramatical de sujeito ou objeto. No arquivo de termos, as informações faltantes são identificadas através do caractere "?", portanto qualquer termo que possua esse caractere entre suas informações contextuais necessárias foi eliminado. A Tabela 7 apresenta a quantidade de termos restantes após a execução da segunda etapa do processo.

Tabela 7. Termos obtidos após a execução da segunda etapa

Domínio	Termos	Eliminação
Pediatria	46.633	75%
Geologia	104.230	76%
Mineração de dados	75.395	69%
Modelagem estocástica	79.343	68%
Processamento paralelo	75.725	69%

Através da Tabela 7, é possível verificar que o percentual de termos eliminados é próximo em todos os *corpus* de domínio. A execução da etapa de eliminação de termos com informações faltantes é essencial para o processo, pois através dela evita-se que o processo seja executado sobre dados incompletos. Do ponto de vista computacional, a diminição de cerca de 70% de dados a serem processados auxilia no aumento de desempenho do sistema computacional.

Após a eliminação dos termos com informações contextuais faltantes, a próxima etapa do processo é a identificação de conceitos. Para isso, o sistema computacional compara os termos restantes com a lista de conceitos de domínio produzida tambem pelo ExATOlp [20]. Nesta etapa,

além da eliminação, para cada um dos termos agora considerados conceitos é obtido o valor de seu índice *tf-dcf* [19], disponibilizado na tabela de conceitos de domínio.

A Tabela 8 apresenta a quantidade de conceitos restantes após as três primeiras etapas do processo, além de apresentar o percetual de eliminção de termos com relação à etapa anterior e ao número inicial de termos.

Tabela 8. Conceitos de domínio após a terceira etapa

Domínio	**Conceitos**	**Eliminação**	**Eliminação Total**
Pediatria	7.410	85%	96%
Geologia	14.728	86%	97%
Mineração de dados	8.078	89%	97%
Modelagem estocástica	11.135	86%	96%
Processamento paralelo	11.431	84%	95%

Conforme pode ser visto na Tabela 8, entre 85% e 89% dos termos obtidos após as duas primeiras etapas do processo não eram conceitos de domínio, e portanto foram descartados. Esta etapa também é fundamental para o processo, pois através dela garante-se que serão extraídas relações não-taxonômicas de conceitos com informações contextuais completas.

A próxima etapa do processo é a extração de relações não-taxonômicas. Nessa etapa, são identificados quais conceitos se relacionam através de um mesmo verbo. Estes conceitos então são classificados de acordo com a sua função sintática (sujeito ou objeto). Após essa classificação, verifica-se quais conceitos se relacionam através de um mesmo verbo em uma mesma sentença. Caso essa situação ocorra, é gerada a tripla definida por <*conceito sujeito, verbo, conceito objeto*>. Para cada uma das triplas também é calculado o índice de Frequência Acumulada e Compartilhada.

A verificação da ocorrência de dois conceitos relacionados pelo mesmo verbo em uma mesma sentença só é possível graças as informações contextuais dos conceitos. Através dos itens 13 e 14 da Tabela 1 (Capítulo 2, Seção 2.2, pg. 22) é possível identificar a frase e o documento em que o conceito está inserido no *corpus* de domínio.

A Tabela 9 apresenta o número de triplas e relações extraídas para cada um dos *corpora* de domínio. Como relações são considerados todos elementos únicos que relacionam os conceitos, no escopo deste trabalho as relações são identificadas pelos verbos únicos obtidos na geração das triplas.

Tabela 9. Triplas e relações não-taxonômicas obtidas

Domínio	Triplas	Relações
Pediatria	159	93
Geologia	306	129
Mineração de dados	128	65
Modelagem estocástica	228	94
Processamento paralelo	374	160

A execução da aplicação do processo não apresentou nenhum problema com relação ao desempenho computacional. Embora o número inicial de dados a serem processados é grande, com as etapas de eliminação de termos a etapa de extração de relações, que exige maior poder de processamento, realizou o processo com um número de dados menor.

Na Figura 12 é possível visualizar relações não-taxonômicas extraídas do *corpus* de domínio de Pediatria. O mapa semântico representa as relações extraídas para os conceitos sujeito “aleitamento_natural”, “aleitamento” e “aleitamento_materno”.

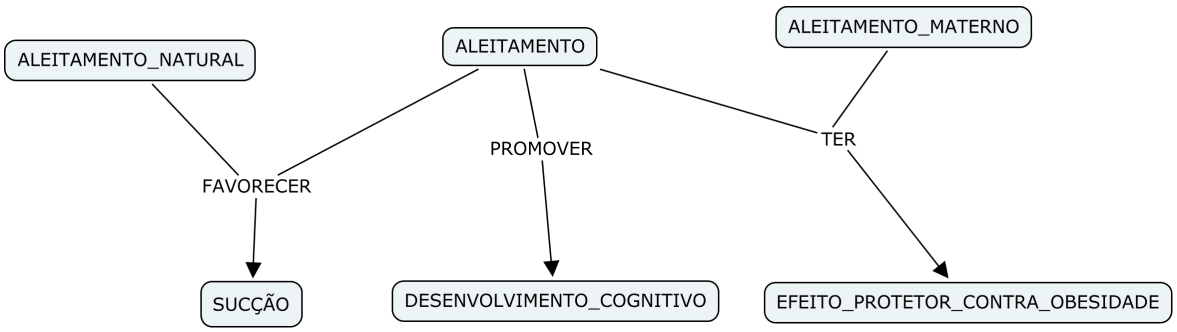

Figura 12. Mapa semântico de relações não-taxonômicas

4.2 Análise

A análise das relações não-taxonômicas extraídas nessa dissertação foi realizada de três formas: (i) análise da especificidade das relações (verbos) extraídas, através da comparação com verbos encontrados no *corpus* do jornal Diário Gaúcho; (ii) análise dos conceitos de domínio identificados como sujeitos ou objetos e (iii) análise das relações não-taxonômicas extraídas do ponto de vista de um especialista no contexto de análise de papéis semânticos.

4.2.1 Análise de Especificidade das Relações

A análise da especificidade das relações extraídas foi realizada por um grupo de três avaliadores, formado por linguistas do curso de Letras da Universidade Federal do Rio Grande do Sul. Entre eles foram divididos 20 verbos extraídos a partir do *corpus* do jornal Diário Gaúcho. Esses verbos foram identificados como os mais utilizados no *corpus* do jornal Diário Gaúcho no mês de junho de 2010. Além disso, os verbos deveriam ocorrer em estruturas trigramas do tipo sujeito+verbo+objeto, sendo necessário que o verbo ocorresse ladeado de sujeito, objeto ou pelos dois.

O padrão dos verbos mais frequentes neste grupo de 3 elementos foi observado e descrito em todo o *corpus*, e sua apresentação foi comparada com o conteúdo do Dicionário Houaiss. Após isso, foram descartados os trigramas que não possuíssem a estrutura sujeito+verbo ou verbo+objeto, verbos de ligação (ser, estar, permanecer, ficar), o verbo haver com sentido de existir e trigramas que se repetiam apenas uma vez.

Os avaliadores verificaram se os verbos selecionados no *corpus* do jornal Diário Gaúcho ocorriam nas relações não-taxonômicas extraídas. Os verbos que ocorriam foram então testados com seu sujeito e objeto com maior valor de índice *tf-dcf* na ferramenta WebCorp [34].

Para cada um dos verbos testados com seu sujeito e objeto mais frequente foi analisado se havia ocorência, caso houvesse foi verificado se ocorria dentro ou fora do domínio de onde a relação foi extraída. Como os verbos extraídos do *corpus* do jornal Diário Gaúcho são cotidianos, a não ocorrência com o sujeito ou objeto, possibilita verificar a especificidade da relação (sujeito+verbo, verbo+objeto).

Por exemplo, verificou-se que o verbo “acompanhar” ocorreu apenas no domínio de Geologia, e com isso foi registrada ocorrência ZERO nos demais domínios. Dentro do domínio de Geologia foi verificado que o sujeito “espinélio” e o objeto “silimanita” são os conceitos com maior índice *tf-dcf* relacionados através do verbo “acompanhar”.

Com isso, foi verificado através da ferramenta WebCorp [34] se o verbo “acompanhar” ocorria com os conceitos “espinélio” e “silimanita” assumindo função sintática tanto de sujeito quanto de objeto. Através desta pesquisa não foi encontrada nenhuma ocorrência, o que reforça a relação entre estes sujeitos e objetos como específica do domínio de Geologia.

Outro verbo selecionado foi “escrever”, que só ocorreu no domínio de Processamento Paralelo. A este verbo estão relacionados o sujeito “ge” e o objeto “mensagem”. Foi então realizada uma pesquisa com a ferramenta WebCorp [34] com esse verbo e esses conceitos. Nesta pesquisa foi encontrada uma ocorrência do conceito “ge” associado ao verbo “escrever”, porém fora do domínio de Processamento Paralelo. Também foram encontradas ocorrências do verbo “escrever” com o conceito “mensagem”, porém em sua maioria fora do domínio técnico, o que permite concluir que esta relação não é específica do domínio.

Através da análise de especificidade, foi possível verificar que a maioria das relações extraídas através do processo proposto por essa dissertação são específicas do domínio. Essa conclusão se dá através da baixa ocorrência dos verbos extraídos do *corpus* do jornal Diário Gaúcho, considerados cotidianos, nas relações não-taxonômicas extraídas, que são consideradas específicas de domínio. Os demais resultados desta análise podem ser visualizados no **Apêndice A** dessa dissertação.

4.2.2 Análise dos Conceitos Extraídos

Além da análise da especificidade das relações (verbos) extraídas, também foram analisados os conceitos identificados como sujeitos e objetos nas relações. Para isso, um grupo de avaliadores formado por alunos do curso de Letras da Universidade Federal do Rio Grande do Sul foi dividido em grupos e cada componente do grupo identificou e classificou os sujeitos e objetos mais relevantes de um domínio.

Para isso, cada avaliador deveria pesquisar informações básicas sobre o domínio que iria analisar. Após, utilizando a ferramenta para visualização de relações não-taxonômicas proposta nessa dissertação, deveria visualizar os sujeitos e objetos do domínio escolhido e ordená-los por maior frequência. Em seguida deveria percorrer toda lista de conceitos e identificar 10 conceitos que ele considerasse bom como sujeito e objeto do domínio.

Para auxiliar na escolha e classificação de 10 sujeitos e objetos bons, o usuário deveria analisar quais foram as relações não-taxonômicas extraídas para o conceito em questão. Além disso, também poderia realizar buscas no Google utilizando o conceito + nome do domínio e na ferramenta WebCorp [34].

Como resultado desse processo, cada avaliador produziu uma lista ordenada por relevância para o domínio contendo 10 conceitos sujeitos e objetos, esse material pode ser visto no **Apêndice B**. A cada um dos termos foi atribuído um peso, seguindo a ordem de classificação. Com isso, o conceito sujeito considerado mais relevante recebeu peso 10, enquanto o 10º conceito mais relevante recebeu peso 1.

As listas de sujeitos e objetos foram agrupadas por domínio e divididas em melhores sujeitos e melhores objetos de um domínio específico. Após, foi contabilizado o peso absoluto de cada conceito de acordo com as classificações de todos avaliadores. Para isso, foi somado o peso para os conceitos que aparecem mais de uma vez nas listas dos avaliadores do domínio. Na Tabela 10, gerada pela soma dos pesos dos conceitos considerados melhores sujeitos para o domínio de Geologia, é possível visualizar um exemplo.

Tabela 10. Melhores sujeitos das relações para o domínio de Geologia

Sujeito	Soma	Automático	Humano
Fácies	86	420	430
Sedimentação	67	423	335
Litologia	54	289,37	270
Bacia	52	359,73	260
Biotita	38	247	190
Litofácies	26	225	130
Granada	24	257	120
Lago	20	278	100
Afloramento	18	218,5	90
Porosidade	16	258	80

A Tabela 10 apresenta os conceitos considerados melhores sujeitos para o domínio de Geologia, a soma dos pesos encontrados para cada conceito selecionado para do referido domínio, o valor índice *tf-dcf* (automático) do conceito e o valor da soma dos pesos nivelado (humano) com o índice *tf-dcf*. Por nivelado, entenda-se que o valor atribuído por humanos foi aproximado aos valores de *tf-dcf* para facilitar a visualização. Nesse exemplo do *corpus* de Geologia o nivelamento correspondeu a multiplicar a soma dos valores atribuídos por humanos por 5 (430 = 86 x 5).

Para cada um dos melhores sujeitos e objetos das relações de um domínio foi gerado um gráfico com o objetivo de comparar o valor da avaliação automática (*tf-dcf*) com o valor da avaliação humana (soma dos pesos dos conceitos definido na classificação pelos avaliadores). Na Figura 13 visualiza-se o gráfico gerado para os melhores sujeitos do domínio de Geologia.

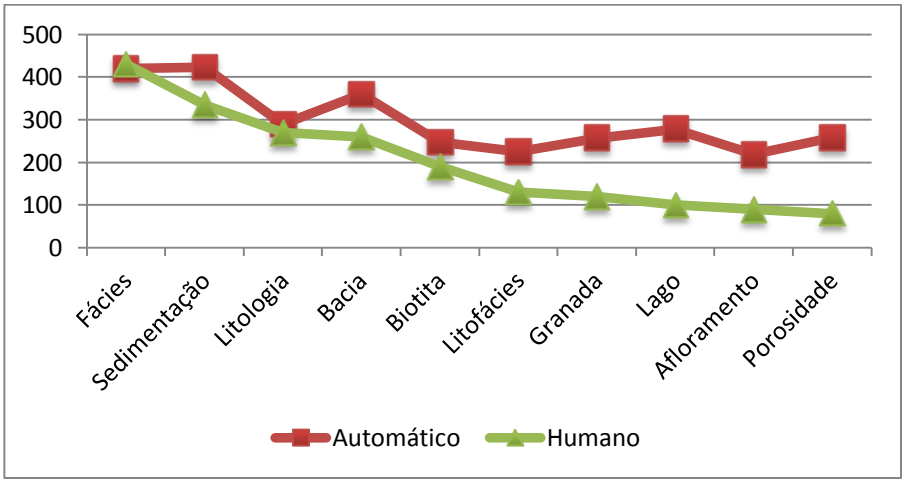

Figura 13. Gráfico dos melhores sujeitos das relações para o domínio de Geologia

Através do gráfico apresentado na Figura 13, verifica-se uma curva similar de correlação forte entre a avaliação humana e a automática. Essa situação permite concluir que a avaliação automática obtida através do índice *tf-dcf* de cada conceito se aproximou da avaliação humana. Também é possível verificar que os conceitos Bacia e Lago apresentam um comportamento diferenciado. Ambos os conceitos relevantes para Geologia foram classificados como melhores sujeitos, porém em níveis de relevância diferente.

Na Figura 14 é possível visualizar o gráfico gerado para os melhores objetos do domínio de Geologia. Neste gráfico verifica-se novamente correlação forte entre a avaliação humana e a avaliação automática. Nesse caso, os conceitos Cimentação e Anfibólio também apresentam comportamento diferenciado.

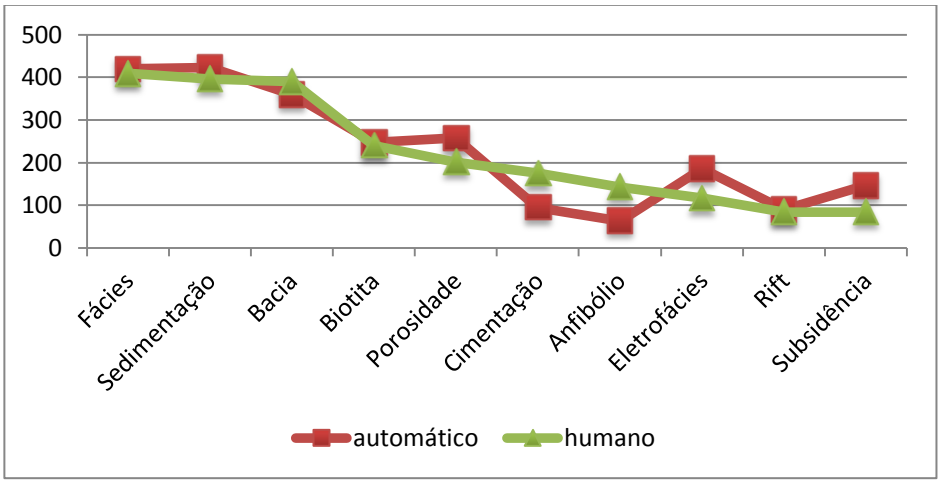

Figura 14. Gráfico dos melhores objetos das relações para o domínio de Geologia

Os mesmo processo foi realizado para o domínio de Pediatria. Na Figura 15 é possível visualizar o gráfico gerado para os melhores sujeitos das relações desse domínio. A curva entre avaliação automática e humana também apresenta correlação forte, com exceção dos conceitos Lactente e IMC. Verifica-se que apenas esses dois conceitos apresentam um valor de *tf-dcf* muito superior.

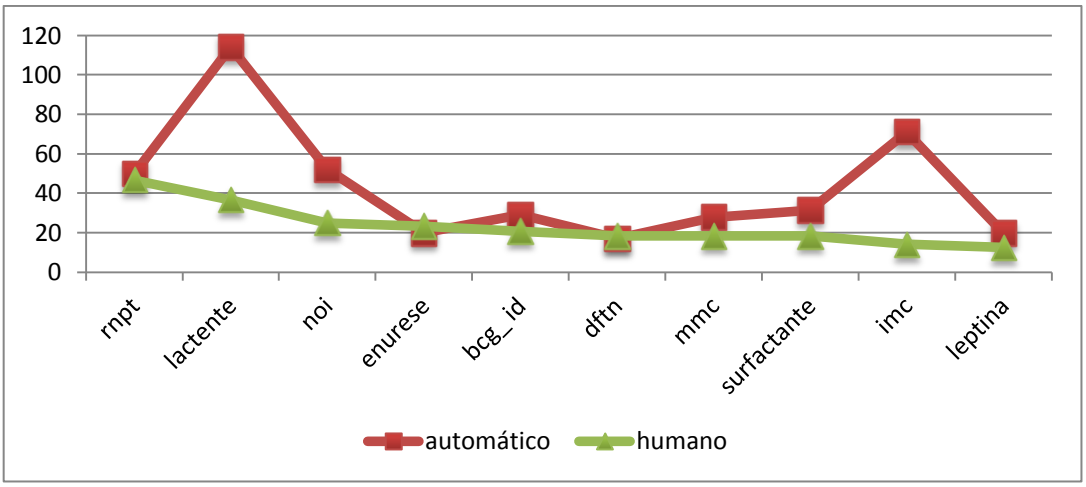

Figura 15. Gráfico dos melhores sujeitos das relações para o domínio de Pediatria

A Figura 16 apresenta o gráfico dos melhores objetos das relações para o domínio de Pediatria. Neste caso, somente os conceitos Lactação, Asma e Hipertensão apresentam comportamente diferenciado. Assim como no gráfico dos melhores sujeitos (Figura 15), apenas estes três conceitos são considerados mais relevantes para Pediatria pelo índice *tf-dcf*.

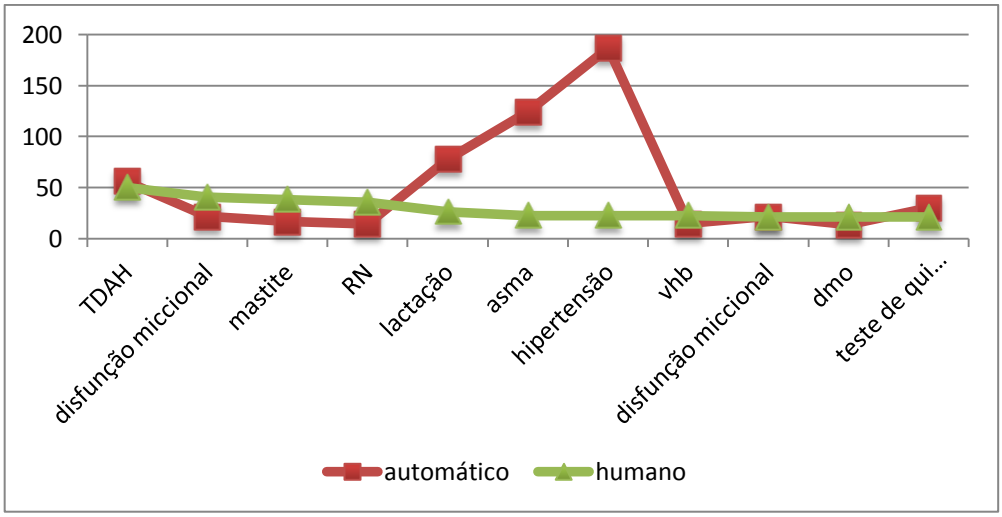

Figura 16. Gráfico dos melhores objetos das relações para o domínio de Pediatria

Na Figura 17 é apresentado o gráfico para os melhores sujeitos do domínio de Mineração de Dados. Assim como nos demais gráficos verifica-se correlação forte entre as curvas da avaliação automática e da avaliação humana. Neste caso, a exceção são os conceitos Contexto Formal e DASE.

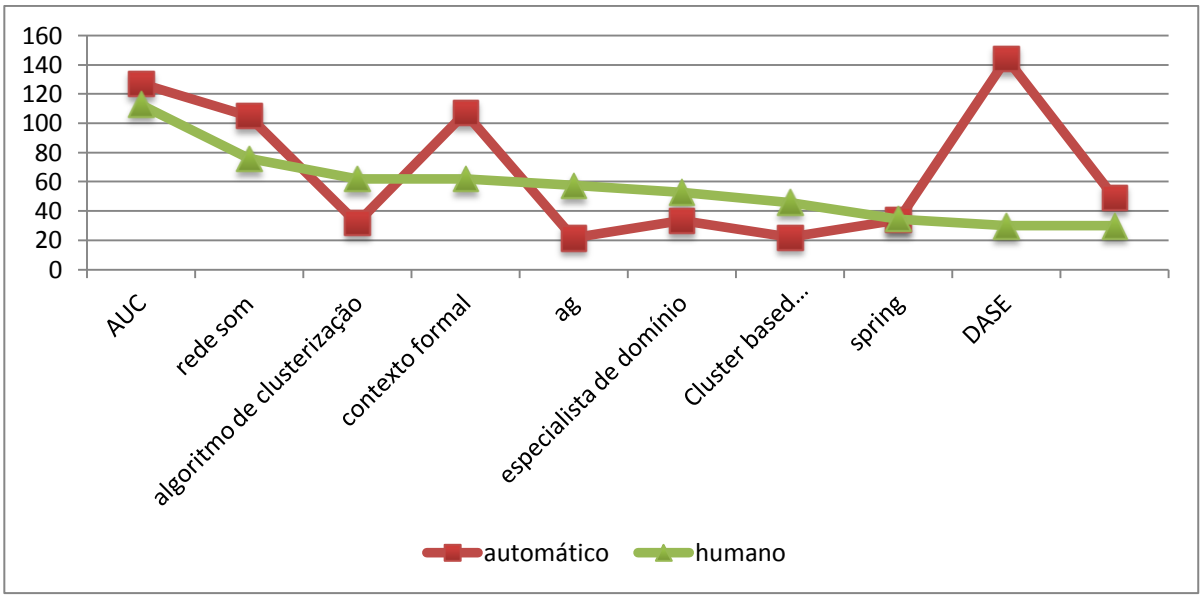

Figura 17. Gráfico dos melhores sujeitos das relações para o domínio de Mineração de Dados

O gráfico dos melhores objetos para o domínio de Mineração de Dados é apresentado na Figura 18. Verifica-se claramente a correlação forte entre a avaliação humana e a avaliação automática. Além do conceito Fonetograma, como na classificação dos melhores sujeitos (Figura 17), o conceito Contexto Formal também apresenta comportamento diferenciado.

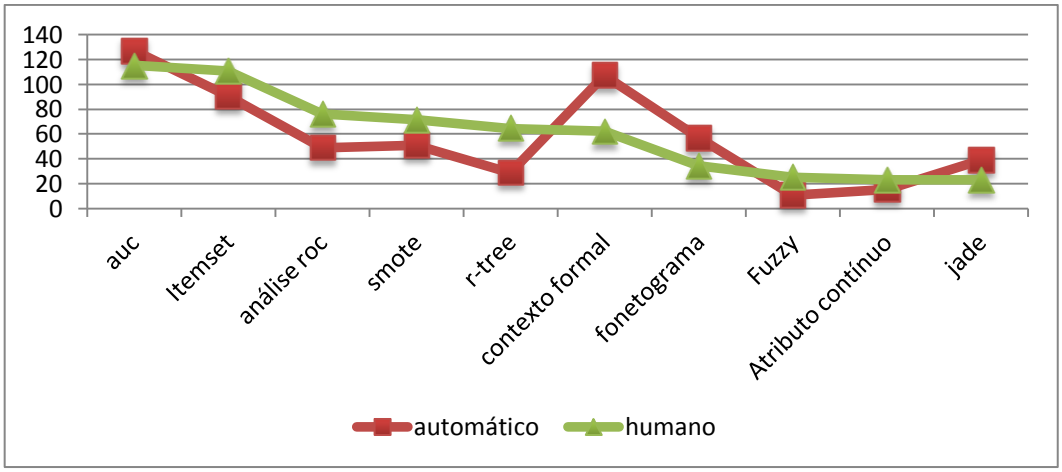

Figura 18. Gráfico dos melhores objetos das relações para o domínio de Mineração de Dados

Os melhores sujeitos das relações para o domínio de Modelagem Estocástica são apresentados no gráfico da Figura 19. Verifica-se neste gráfico correlação forte entre a avaliação

humana e automática dos conceitos. Neste caso, os conceitos Modelo SAN e Servidor RIO apresentam comportamento diferenciado.

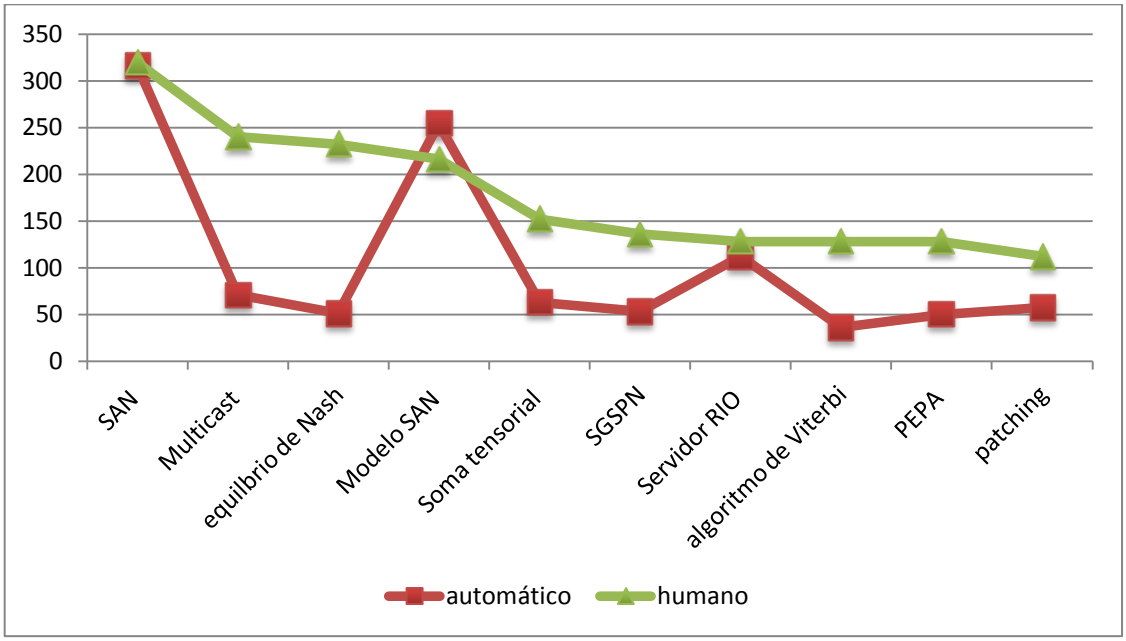

Figura 19. Gráfico dos melhores sujeitos das relações para o domínio de Modelagem Estocástica

Assim como nos melhores sujeitos para o domínio de Modelagem Estocástica (Figura 19), para os melhores objetos o conceito Modelo SAN apresenta comportamento diferenciado entre as avaliações humana e automática (Figura 20). Isso ocorre pelo fato de que esse conceito não é um objeto do domínio, e sim sujeito, tendo sido classificado de forma errônea. Além de Modelo SAN, os conceitos SAN e HMM também apresentam comportamento diferenciado.

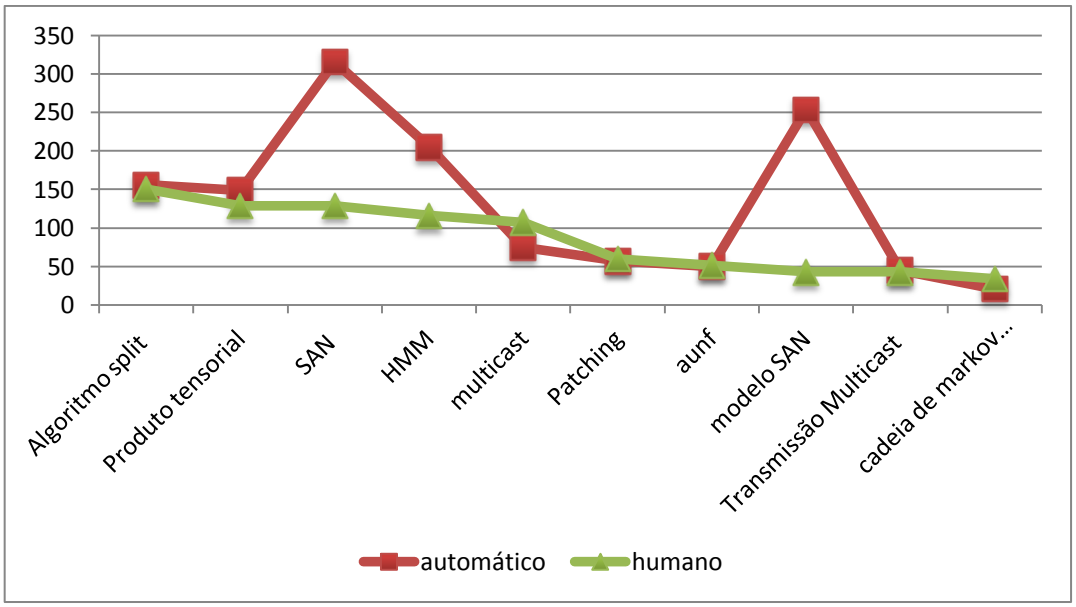

Figura 20. Gráfico dos melhores objetos das relações para o domínio de Modelagem Estocástica

Os melhores sujeitos para o domínio de Processamento Paralelo são apresentados na Figura 21. Neste caso a curva entre a avaliação humana e automática não apresenta grandes variações. Embora o conceito Checkpoint apresente uma avaliação automática muito superior aos demais, ele foi classificado como mais relevante também na avaliação humana.

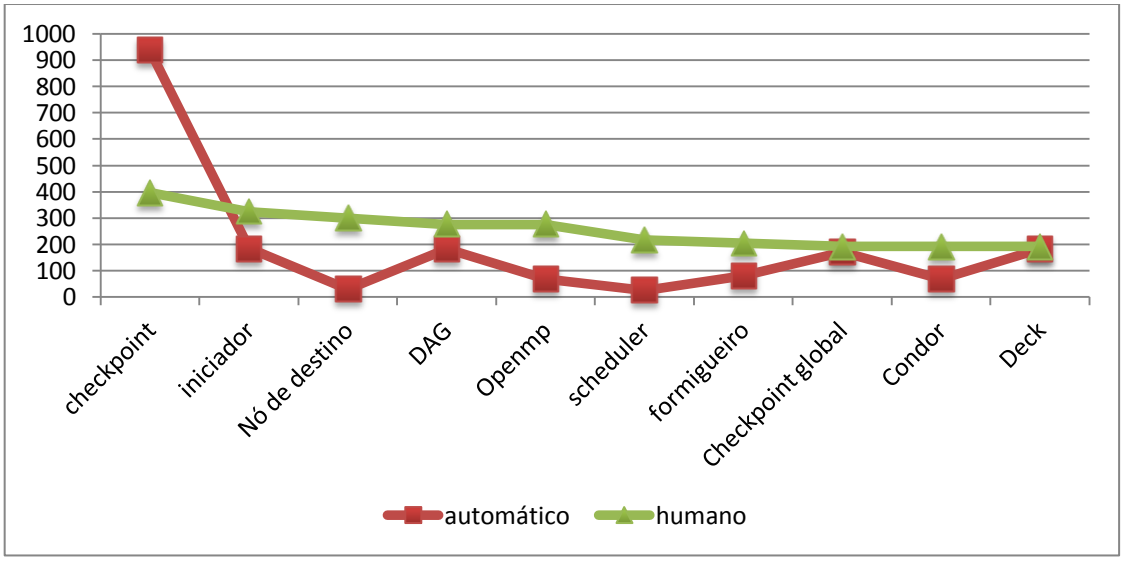

Figura 21. Gráfico dos melhores sujeitos das relações para o domínio de Processamento Paralelo

Assim como para os melhores sujeitos (Figura 21), na classificação dos melhores objetos para o domínio de Processamento Paralelo não houve grande variação entre as curvas de avaliação (Figura 22). Nesse caso apenas o conceito checkpoint apresentou comportamento diferenciado.

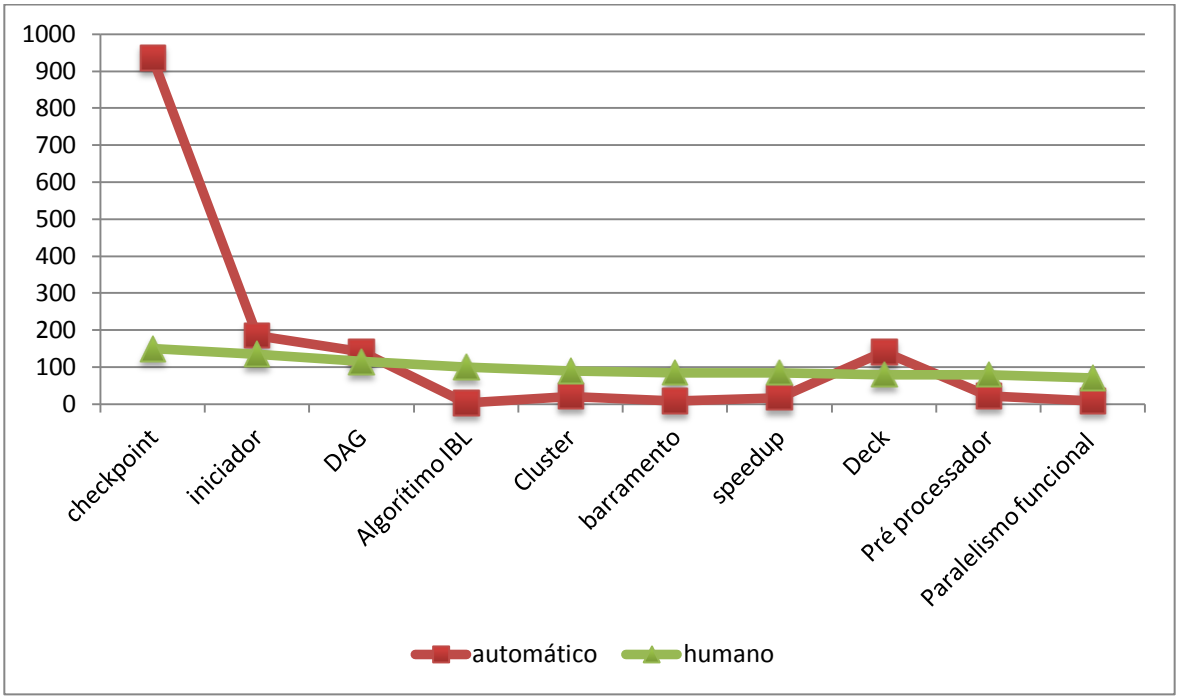

Figura 22. Gráfico dos melhores objetos das relações para o domínio de Processamento Paralelo

4.2.3 Análise da Aplicação das Relações Extraídas

Com o objetivo de analisar as relações não-taxonômicas extraídas, um avaliador, doutorando em Estudos da Linguagem do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul, relatou sua utilização no contexto da análise de papéis semânticos.

O objetivo do trabalho do avaliador é anotar um *corpus* especializado e um *corpus* não especializado com papéis semânticos para realizar a comparação entre os dois *corpora*. As sentenças anotadas dos dois *corpora* serão disponibilizadas (XML) após a anotação para que possam ser utilizadas por outros pesquisadores. Após a anotação, será feita uma comparação com a VerbNet em inglês para observar os potenciais dos papéis semânticos para os estudos de tradução automática.

De acordo com o avaliador a ferramenta para visualização das relações é muito útil para a observação de colocações verbais. Ainda segundo o avaliador, as relações apresentadas, quando inseridas em uma ontologia, auxiliam na compreensão das relações entre termos das areas estudadas e possivelmente da linguagem comum também. Além disso, com um *corpus* paralelo ou comparável as relações auxiliariam muito na construção de um dicionário bilíngüe de colocações.

O avaliador também destacou dois fatores negativos das relações extraídas quando utilizadas no contexto da anotação de papéis semânticos. O primeiro é que as relações não apresentam todos os complementos do verbo, pois alguns verbos tem mais complementos do que apenas o sujeito e o objeto. Como exemplo o avaliador citou a relação entre os conceitos "Harris" e "Ausência de Biotita" através do verbo "associar" no domínio de Geologia. Nesse caso, de acordo com o avaliador, "Harris" foi quem descobriu ou fez a associação, mas a associação em si é entre dois elementos e um deles não está presente. O outro ponto negativo descatado pelo avaliador é que não são apresentadas as preposições que ligam os complementos, no caso dos verbos transitivos indiretos, e nem adjetivos.

5 CONSIDERAÇÕES FINAIS

Neste Capítulo serão apresentadas as contribuições e conclusões obtidas nessa dissertação. Além disso, também serão apresentados trabalhos futuros que venham complementar o processo proposto ou que possam fazer uso das relações não-taxonômicas extraídas.

5.1 Contribuições e Conclusões

O objetivo central dessa dissertação foi a proposta de um processo de extração automática de relações não-taxonômicas a partir de *corpus* em língua portuguesa. Para isso, diferente dos trabalhos encontrados na literatura, foram utilizados como fonte de dados conceitos de domínio e informações contextuais extraídas a partir de *corpus* em língua portuguesa através da ferramenta ExATOlp [20]. Como produto do processo proposto foram extraídas de forma automática relações não-taxonômicas e o seu valor de Frequência Acumulada e Frequência Compartilhada obtidos através do índice *tf-dcf* de cada conceito.

O objetivo foi alcançado e o processo foi aplicado sobre cinco *corpora* de domínio. Sobre os resultados da aplicação foram realizadas três avaliações que tinham como objetivo verificar a especificidade das relações em relação ao domínio, a relevância dos conceitos sujeitos e objetos para o domínio em questão e as relações não-taxonômicas extraídas do ponto de vista de um especialista.

Conforme pode ser visto através da análise de especificidade das relações (verbos) extraídas, o processo proposto nessa dissertação extraiu relações pertinentes ao domínio em questão. Assim como também pode ser visto através da análise de relevância dos conceitos considerados sujeitos e objetos de domínio que os sujeitos e objetos extraídos representam informações relevantes para o domínio.

Na análise da relevância dos sujeitos e objetos, embora tenha ocorrido de alguns conceitos distoarem dos demais, houve correlação forte entre a avaliação humana e a avaliação automática (*tf-dcf*) dos conceitos. Sendo assim pode-se destacar como contribuição o fato de que o índice *tf-dcf* é uma medida válida para estimar a relevância de uma relação não-taxonômica, assim como os conceitos com informações contextuais também são úteis para extração de relações.

Através da análise do especialista sobre a utilização das relações não-taxonômicas no contexto de anotação de papéis semânticos, foi possível concluir que as relações extraídas podem ser úteis em aplicações linguísticas. Dentre estas aplicações pode-se citar a construção de um dicionário bilíngue ou um recurso para compreensão de relações entre termos específicos de um domínio.

O foco dessa dissertação é identificar apenas relações não-taxonômicas entre conceitos. Com isso, as relações extraídas não apresentam complementos para os verbos além do sujeito e objeto relacionado, preposições ou adjetivos. Embora, nas informações contextuais dos conceitos também seja possível identificar adjetivos. Com isso, pode-se concluir que foi proposto um processo para extração automática de relações não-taxonômicas relevante

Dentre as contribuições dessa dissertação pode-se citar o processo para extração automática de relações não-taxonômicas em *corpus* da língua portuguesa proposto. Conforme já citado, não há nenhum trabalho que apresente uma proposta para a língua portuguesa. Ainda nesse contexto,

nenhuma das propostas encontradas na literatura utiliza como fonte de dados os conceitos de um domínio com suas informações contextuais.

Ainda como contribuição dessa dissertação, pode ser citada a ferramenta para visualização de relações não-taxonômicas proposta. Através dela foi possível desenvolver as três atividades de avaliação, sem necessidade de interferência com relação a esclarecimentos sobre o seu uso.

5.2 Trabalhos Futuros

Os trabalhos futuros dessa dissertação podem ser divididos em três eixos de pesquisa: (i) aplicação das relações não-taxonômicas extraídas; (ii) aprimoramento do processo de extração; e (iii) construção automática de ontologias.

Dentro do eixo de aplicação das relações não-taxonômias extraídas, podem ser exploradas possibilidades de utilização das relações em contextos variados. Dentre eles encontra-se aplicação em testes para avaliar a compreensão de termos técnicos na língua portuguesa, a construção automática de um dicionário de conceitos de domínio, que apresenta informações básicas sobre os conceitos e o desenvolvimento de um concordanciador com base nas relações extraídas.

No eixo de aprimoramento do processo de extração destaca-se a possibilidade de aplicação de um filtro maior sobre as relações não-taxonômicas extraídas. Por exemplo, como a eliminação de relações que não são específicas do domínio, ou de relações identificads por verbos de ligação. Para que isso seja possível, um estudo maior sobre a classificação de relevância das relações deve ser desenvolvido também. Além disso também traria contribuições a análise estatística sobre a correlação da curva de análise humana e automática sobre os conceitos extraídos como sujeitos e objetos.

Extendendo ainda mais os resultados do processo de extração de relações não-taxonômicas proposto nessa dissertação também é possível trabalhar no eixo de construção automática de ontologias. Visto que o processo de extração de relações não-taxonômicas é negligenciado na construção de ontologias, pode-se desenvolver um módulo para o aplicativo ExATOlp que inclua em uma ontologia as relações não-taxonômicas extraídas pelo processo dessa dissertação.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Berners-Lee, T.; Hendler, J.; Lassila, O. "The semantic web". *Scientic American*, vol. 284, 2001, pp. 34-43.

[2] Biemann, C. "Ontology learning from text: A survey of methods". *LDV Forum*, vol. 20, 2005, pp. 75-93.

[3] Brewster, C.; Ciravegna, F.; Wilks, Y. "Background and foreground knowledge in dynamic ontology construction". In: SIGIR Semantic Web Workshop, 2003, 8p.

[4] Brewster, C.; Jupp, S.; Luciano, J.; Shotton, D.; Stevens, R. D.; Zhang, Z. "Issues in learning an ontology from text". *BMC Bioinformatics*, vol. 10, 2009, pp. 1-20.

[5] Buitelaar, P.; Cimiano, P.; Magnini, B. "Ontology learning from text: An overview". *Ontology Learning from Text: Methods, Evaluation and Applications, Paul Buitelaar, Philipp Cimiano, e Bernardo Magnini (Eds)*, vol. 123, 2005, pp. 3-12.

[6] Byrd, R.; Ravin, Y. "Identifying and extracting relations from text". In: 4th International Conference on Applications of Natural Language to Information Systems, 1999, 5p.

[7] Chung, T. M. "A corpus comparison approach for terminology extraction". *Terminology*, vol. 9, 2003, pp. 221-246.

[8] Cimiano, P.; Volker, J.; Studer, R. "Ontologies on demand? – A description of the state-of-the-art, applications, challenges and trends for ontology learning from text". *Information, Wissenschaft und Praxis*, vol. 57, 2006, pp. 315-320.

[9] Drouin, P. "Detection of domain specific terminology using corpora comparison". In: 4th International Conference on Language Resources and Evaluation, 2004, pp. 79-82.

[10] Faure, D.; Nedellec, C. "A corpus-based conceptual clustering method for verb frames and ontology acquisition". In: LREC Workshop on Adapting Lexical and Corpus Resources to Sublanguages and Applications, 1998, pp. 5-12 .

[11] Ferreira, V. H.; Lopes, L.; Vieira, R. "Descoberta automática de relaçõs não-taxonômicas a partir de corpus em língua portuguesa". In: ONTOBRAS/MOST 5th Seminar on Ontology Research in Brazil and 7th International Workshop on Metamodels, Ontologies and Semantics Technologies, 2012, pp. 1-6

[12] Finkelstein-Landau, M.; Morin, E. "Extracting semantic relationships between terms: supervised vs. unsupervised methods". In: International Workshop on Ontological Engineering on the Global Information Infrastructure, 1999, pp. 71-80.

[13] Gruber, T. "Toward principles for the design of ontologies used for knowledge sharing". *International Journal Human-Computer Studies*, vol. 43, 1993, pp. 907-928.

[14] Guarino, N.; Schneider, L. "Ontology-driven conceptual modeling". *Lecture Notes in Computer Science*, vol. 2503, 2002, 10p.

[15] Hahn, U.; Schnattinger, K. "Towards text knowledge engineering". In: Association for the Advancement of Artificial Intelligence, 1998, pp. 524–531.

[16] Kavalec, M.; Maedche, A.; Svátek, V. "Discovery of lexical entries for non-taxonomic relations in ontology learning". *Lecture Notes in Computer Science*, vol. 2932, 2004, pp. 249–256

[17] Kavalec, M.; Svátek, V. "A study on automated relation labelling in ontology learning". *Ontology Learning from Text: Methods, Evaluation and Applications, Paul Buitelaar, Philipp Cimiano, e Bernardo Magnini (Eds)*, 2005, pp. 44–58.

[18] Kim, S. N.; Baldwin, T.; Kan, M. Y. "Extracting domain-specic words: A statistical approach". In: Australasian Language Technology Association Workshop, 2009, pp. 94-98.

[19] Lopes, L. "Extração Automática de Conceitos a Partir de Textos na Língua Portuguesa", Tese de Doutorado, Programa de Pós-Graduação em Ciência da Computação, PUCRS, 2012, 156p.

[20] Lopes, L.; Fernandes, P.; Vieira, R.; Fedrizzi, G. "ExATOlp - An automatic tool for term extraction from portuguese language corpora". In: 4th Language and Technology Conference, 2009, pp. 427-431.

[21] Maedche, A.; Staab, S. "Mining non-taxonomic conceptual relations from text". In: 12th European Knowledge Acquisition Workshop, 2000, pp. 2-6.

[22] Milios, E; Zhang, Y.; Dong, L. "Automatic term extraction and document similarity in special text corpora". In: 6th Conference of the Pacific Association for Computational Linguistics, 2003, pp. 275-284.

[23] Morin, E. "Automatic acquisition of semantic relations between terms from technical corpora". In: 5th International Congress on Terminology and Knowledge Engineering, 1999, 10p.

[24] Nabila, N. F.; Mamat, A.; Azmi-Murad, M. A.; Mustapha, N. "Enriching non-taxonomic relations extracted from domain texts". In: International Conference on Semantic Technology and Information Retrieval, 2011, pp. 99-105.

[25] Noy, N. F.; McGuinness D. L. "What is an Ontology and why we need it". Capturado em: http://liris.cnrs.fr/alain.mille/enseignements/Ecole_Centrale/What%20is%20an%20ontology%20and%20why%20we%20need%20it.htm, Outubro 2012.

[26] Pantel, P.; Lin, D. "A statistical corpus-based term extractor". In: 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Articial Intelligence, 2001, pp. 36-46.

[27] Park, Y.; Patwardhan, S.; Visweswariah, K.; Gates, S. C. "An empirical analysis of word error rate and keyword error rate". In: INTERSPEECH, 2008, pp. 2070-2073.

[28] Salton, G.; Buckley, C. "Term weighting approaches in automatic text retrieval". *Information Processing and Management*, vol. 24, 1988, pp. 513–523.

[29] Sanchéz, D.; Moreno, A. "Learning non-taxonomic relationships from web documents for domain ontology construction". *Data and Knowledge Engineering*, vol. 64, 2008, pp. 600-623.

[30] Schutz, A.; Buitelaar, P. "RelExt: A tool for relation extraction in ontology extension". In: 4th International Semantic Web Conference, 2005, pp. 593–606.

[31] Serra, I.; Girardi, R. "A process for extracting non-taxonomic relationship of ontologies from text". *Intelligent Information Management*, vol. 3, 2011, pp. 119-124.

[32] Swartout, B. "Toward distributed use of large-scale ontologies". Capturado em: http://ksi.cpsc.ucalgary.ca/kaw/kaw96/swartout/Banff_96_final_2. html, Maio 2012.

[33] Villaverde, J.; Persson, A.; Godoy, D.; Amandi, A. "Supporting the discovery and labeling of non-taxonomic relationships in ontology learning". *Experts Systems with Applications*, vol. 36, 2009, pp. 10288-10294.

[34] WEBCORP. Capturado em: http://www.webcorp.org.uk, Dezembro 2012.

[35] Weichselbraun, A.; Wohlgenannt, G.; Scharl, A.; Granitzer, M.; Neidhart, T.; Juffinger, A. "Discovery and evaluation of non-taxonomic relations in domain ontologies". *International Journal of Metadata, Semantics and Ontologies*, vol. 4, 2009, pp. 212-222.

[36] Lopes, L.; Fernandes, P.; Vieira, R. "Domain term relevance through tf-dcf". In: ICAI – International Conference in Artificial Inteligence, 2012, pp. 1-7.

A AVALIAÇÃO DE ESPECIFICIDADE

Verbo	Pediatria	Geologia	Mineração de Dados	Modelagem Estocástica	Processamento Paralelo
Acompanhar		35,00 No webcorp não foram encontradas ocorrências de ESPINÉLIO ACOMPANHAR (acompanhou, acompanha, acompanhava) , nem de ACOMPANHAR (acompanhou, acompanha, acompanhava) (a)SILLIMANITA			
Acontecer	ZERO	ZERO	ZERO	ZERO	ZERO
Ajudar	ZERO	ZERO	ZERO	ZERO	ZERO
Assistir	ZERO	ZERO	ZERO	ZERO	ZERO
Cobrar	ZERO	ZERO	ZERO	ZERO	ZERO
Conhecer	ZERO	ZERO	ZERO	ZERO	ZERO
Entrar	ZERO	ZERO	ZERO	ZERO	ZERO
Enviar	ZERO	ZERO	ZERO	88,18	7.000 No webcorp não foi encontrada nenhuma ocorrência do grupo JPVMD ENVIAR (ENVIA, verbo buscado apenas nessa forma). Já com ENVIAR MENSAGEM (ENVIA) foi encontrado um número altíssimo de ocorrências, de modo a não ser

					possível verificar o domínio.
Escolher	ZERO	ZERO	ZERO	ZERO	ZERO
Escrever	ZERO	ZERO	ZERO	ZERO	12, 43 No webcorp, encontrada apenas 01 ocorrência do grupo GE ESCREVER (ESCREVE) e o texto não pertence ao domínio processamento paralelo. Já com ESCREVER MENSAGEM (ESCREVE) foi encontrado um número altíssimo de ocorrências, de modo a não ser possível verificar o domínio.
Estrear	ZERO	ZERO	ZERO	ZERO	ZERO
Falar	ZERO	ZERO	ZERO	ZERO	ZERO
Fazer	ZERO	124	ZERO	216	4.000 No webcorp, encontrada apenas 01 ocorrência do grupo CHECKPOINT FAZER (FAZ, verbo buscado apenas nesta forma) e o texto não pertence ao domínio processamento paralalelo. Também ocorre 01 vez o grupo

					FAZER CKECKPOINT, desta vez no domínio.
Gostar	ZERO	ZERO	ZERO	ZERO	ZERO
Ir	ZERO	ZERO	20,66 No webcorp não há resultados para o sujeito mais frequente AGENTE SCHEDULER nem para o objeto TRANSAÇÃO.	ZERO	ZERO
Ocorrer	53,00 No webcorp não há resultados para o sujeito mais frequente FENÔMENO DE RAYNAUD, nem para o objeto MAMILO.	45,00	8,99	ZERO	ZERO
Querer	ZERO	ZERO	ZERO	ZERO	ZERO
Saber	ZERO	ZERO	ZERO	ZERO	ZERO
Ter	1.498,45 No webcorp,11 ocorrências para ALEITAMENTO MATERNO como sujeito do verbo ter, todos no domínio. Foram encontradas 86 ocorrências para FEBRE como objeto e 81 estavam no domínio.	1.097,25	208,82	546,78	206,16
Ver	165,00 No webcorp foram encontradas 2 ocorrências para VER LEITE MATERNO, mas nenhuma no domínio.	ZERO	ZERO	68,00	ZERO

B AVALIAÇÃO DE RELEVÂNCIA DOS CONCEITOS

Domínio	Geologia
Grupo	1
Sujeito	Objeto
sedimentação	fácies
fácies	biotita
litofácies	rift
litologia	dolomita
biotita	pedogênese
plagioclásio	gradiente de relevo
hangingwall	empilhamento estratigráfico
caolinita	pleocroísmo
eustasia	sedimentação
metamorfismo	anfibólio

Domínio	Pediatria
Grupo	1
Sujeito	Objeto
Rnpt	Disfunção Miccional
Enurese	TDAH
Leptina	dmo
Dftn	vhb
Noi	Duplo Cego
AAP	Mastite
Surfactante	Hipóxia
Mmc	Fibra Insolúvel
DSM IV	Colite
BCG PC	Uso de Fenoterol

Domínio	Processamento Paralelo
Grupo	1
Sujeito	Objeto
Shivaratri	Barramento
Ferramenta de shell	Matriz de rigidez
Sistema solaris	Cliente de grade
Primitiva	Processo mestre
Função mpi	Algoritmo ibl
Nó de destino	Escalonamento round robin
Modelo assíncrono	Modelo síncrono
Yasmin	Espaço de endereçamento
Arquitetura superescalar	Driver de dispositivo
Checkpoint provisório	Ambiente paralelo

Domínio	Modelagem Estocástica
Grupo	1
Sujeito	Objeto
SAN	modelo SAN
SGSPN	HMM
equilbrio de Nash	Produto tensorial
PEPA	multicast
algoritmo de Viterbi	unicast
árvore de atingibilidade	patching
bittorrent	NACK
biblioteca rml	modelo de Gilbert
ACK	tempo discreto
cadeia de morte	descritor Kronecker

Domínio	Mineração de Dados
Grupo	1
Sujeito	Objeto
Ag	Resultado de mineração
Algoritmo de clusterização	Análise Roc
Cluster based oversampling	Atributo contínuo
Miningserver	R tree
Ferramenta visual data minner	Itemset
Sentença SQL	Smote
Lhs	Driver JDBC
Ar frio	Análise de Pareto
Spring	Análise espacial
Tomek	Galois

Domínio	Processamento Paralelo
Grupo	2
Sujeito	Objeto
checkpoint	checkpoint
iniciador	iniciador
formigueiro	máquina_virtual
monitor_central	espaço_de_endereçamento
núcleo_de_comunicação	escalonador
mensagem_de_liberação	valor_de_semáforo
nó_de_destino	paralelismo_funcional
protocolo_rdt	barramento
scheduler	escalonamento
kernel	speedup

Domínio	Pediatria
Grupo	2
Sujeito	Objeto
ascaridíase	teste_de_qui_quadrado
lactente	mastite
manitol	disfunção_miccional
enurese	lactação
leptina	otite
anemia_falciforme	asma
esofagite	lesão
gestação	morbidade
aleitamento	mama
amamentação	amamentação

Domínio	Mineração de Dados
Grupo	2
Sujeito	Objeto
AUC	AUC
contexto formal	contexto formal
rede som	Itemset
ferramenta visual dataminer	fonetograma
spring	smote
especialista de domínio	análise ROC
algoritmo de clusterização	antecedente de regra
agente daseservicelocator	JADE
cluster based oversampling	R-Tree
ag	processo de poda

Domínio	Geologia
Grupo	2
Sujeito	Objeto
Sedimentação	aleitamento_gradacional
Fácies	anfibólio
Litologia	Arcabouço_bioestratigráfico
Biotita	argilosidade
Litofácies	Balanço_hidrico_negativo
Afloramento	Biotita_monogranizo
Planície	Biotita_monogranizo_porfiritrico
Quartzo	Contato_basal
Sistema_Fluvial	Contato_basal_abrupto
Magmatismo	Corrente_de_turbidez_hiperpicnal

Domínio	Mineração de Dados
Grupo	3
Sujeito	Objeto
AUC	AUC
Contexto formal	Contexto formal
Rede SOM	Itemset
Especialista de domínio	Fonetograma
Algoritmo de clusterização	SMOTE
Cluster based oversampling	Análise ROC
AG	Antecedente de regra
Condição de junção	R-tree
Subárvore	Impec
Aprem-IR	Peso binário

Domínio	Pediatria
Grupo	3
Sujeito	Objeto
Amamentação	Bebê
Bebê	RN
RNPT	Amamentação
Gestação	Déficit de crescimento
Enurese	Asma
MMC	Leite materno
BCG-ID	Obesidade
DFTN	TDAH
Uso antenatal de corticosteróides	Mastite
Ausência de aleitamento	Desenvolvimento de DBP

Domínio	Geologia
Grupo	3
Sujeito	Objeto
Fácies	Sedimentação
Cristal	Bacia
Diagênese	Porosidade
Intemperismo	Biotita
Granito	Cimentação
Caladão	*Rift*
Muscovita	Soerguimento
Greenstone	Anfibólio
Veio	freático
LMzMt	Estratificação

Domínio	Modelagem Estocástica
Grupo	3
Sujeito	Objeto
SAN	SAN
HMM	HMM
Algoritmo split	Algoritmo split
Produto tensorial	Produto tensorial
Multicast	Multicast
Patching	Patching
Transmissão Multicast	Transmissão Multicast
NACKs	NACKs
Volatilidade	Volatilidade
Matriz de termo	Matriz de termo

Domínio	Processamento Paralelo
Grupo	3
Sujeito	Objeto
Checkpoint	Checkpoint
Iniciador	Iniciador
Deck	Deck
DAG	DAG
Virtualização	Virtualização
Cluster	Cluster
Ambiente paralelo	Ambiente paralelo
Matriz de rigidez	Matriz de rigidez
Speedup	Speedup
Escalonador	Escalonador

Domínio	Geologia
Grupo	4
Sujeito	Objeto
sedimentação	sedimentação
fácies	Fácies
bacia	Bacia
afloramento	eletrofácies
planície	lençol freático
sistema_fluvial	ângulo_de_contato
magmatismo	recristalização
metamorfismo	geminação
simulador_de_fluxo	espaço_poroso
lâmina	tensão_superficial

Domínio	Mineração de Dados
Grupo	4
Sujeito	Objeto
AUC	AUC
Rede som	Itemset
Apriori	Análise ROC
Spring	JADE
Ag	SMOTE
Lhs	r-tree
Especialista de domínio	Shapefile
DASE	Fuzzy
Algoritmo de clusterização	Análise de Pareto
Plataforma de agentes	Fonetograma

Domínio	Processamento Paralelo
Grupo	4
Sujeito	Objeto
cluster	cluster
mpi	mpi
escalonamento	escalabilidade
kernel	loop
pvm	speedup
openmp	aplicação paralela
daemon	ambiente paralelo
scheduler	corba
benchmarking	cardinalidade
Plataforma de agentes	lei de amdahl

Domínio	Geologia
Grupo	5
Sujeito	Objeto
Sedimentação	Sedimentação
Fácies	Fácies
Bacia	Bacia
Litologia	Porosidade
Lago	Biotita
Porosidade	Eletrofácies
Granada	Subsidência
Biotita	Nível de base
Litofácies	Cimentação
Afloramento	Soerguimento

Domínio	Modelagem Estocástica
Grupo	5
Sujeito	Objeto
SAN	SAN
Modelo San	Algoritmo Split
Servidor Rio	Produto Tensorial
Multicast	Recompensa
Soma tensorial	Multicast
Equilíbrio de Nash	Controle de admissã
PEPA	Transmissão Multicast
Algoritmo de Viterbi	Volatilidade
Árvore de atingibilidade	Temporizador
Taxa de disparo	Distribuição Condicional

Domínio	Processamento Paralelo
Grupo	5
Sujeito	Objeto
checkpoint	checkpoint
iniciador	iniciador
deck	deck
dag	checkpoint global
checkpoint global	dag
condor	checkpoint mutável
openmp	checkpoint provisório
UP	Ambiente paralelo
Lei de Amdahl	lei de amdahl
MPI	cluster

Domínio	Mineração de Dados
Grupo	5
Sujeito	Objeto
auc	auc
contexto_formal	contexto_formal
rede_som	itemset
especialista_de_domínio	smote
algoritmo_de_clusterização	análise_roc
cluster_based_oversampling	intensão
ag	r_tree
condição_de_junção	processo_de_poda
subárvore	atributo_contínuo
atributo_contínuo	peso_binário

Domínio	Pediatria
Grupo	5
Sujeito	Objeto
lactente	hipertensão
noi	tdah
rnpt	teste_de_qui_quadrado
imc	disfunção_miccional
vídeo_eeg	mastite
esofagite	vhb
surfactante	rn
bcg_id	dmo
enurese	pic
leptina	hic

Domínio	Mineração de Dados
Grupo	6
Sujeito	Objeto
Dase	Auc
Auc	Itemset
Agente daseservicelocator	Fuzzy
Aprem ir	R tree
Agente Scheduler	Cluster
Milprit	Shapefile
Pln	Dab
Acsa	Df
Mvc	Agente concurrecycontroller
gspl	Smote

Domínio	Geologia
Grupo	6
Sujeito	Objeto
Litologia	Acomodação
Biotita	Anfibólio
Sedimentação	Bacia
Granada	Basculamento
Fácies	Cimentação
Litofácies	Compactação
Afloramento	Deposição
Variograma	Estratificação
Magmatismo	Guiana
Litologia	Acomodação

Domínio	Processamento Paralelo
Grupo	6
Sujeito	Objeto
Scheduler	Scheduler
Lei de Amdahl	Lei de Amdahl
Balanceamento de Carga	Tamanho de Problema
Carga de trabalho	Pré-processador
Virtualização	Virtualização
Nó de destino	Algorítimo IBL
Nó de origem	Zero-Copy
Checkpoint	Checkpoint
Servidor	Autenticação
Atacante	Certificado Digital

Domínio	Pediatria
Grupo	6
Sujeito	Objeto
oms	rn
rn	tdah
rnpt	disfunção miccional
mmc	desenvolvimento cognitivo
aleitamento	ultra sonografia
lactente	lactação
anemia falciforme	intubação
surfactante	prevalência
dftn	morbidade
gestação	amamentação

Domínio	Modelagem Estocástica
Grupo	6
Sujeito	Objeto
equilíbrio de nash	algoritmo split
patching	aunf
multicast	cadeia de markov equivalente
algoritmo de viterbi	descritor kronecker
número de aunfs	riommclient
árvore de atingibilidade	modelo de gilbert
taxa de disparo	caminhada de lévy
gtaexpress	ataque especulativo
solução shuffle	volatilidade
pepa	janela de transmissão

Domínio	Geologia
Grupo	7
Sujeito	Objeto
sedimentação	sedimentação
fácies	fácies
bacia	porosidade
litologia	biotita
lago	feição
porosidade	planície
granada	eletrofácies
biotita	subsidiência
litofácies	nível de base
afloramento	cimentação

Domínio	Pediatria
Grupo	7
Sujeito	Objeto
lactente	obesidade
imc	hipertensão
noi	asma
rnpt	morbidade
vídeo_eeg	síndrome
biópsia	morbidade
esofagite	lactação
surfactante	intubação
bcg_ id	tdah
endoscopia	sintomatologia

Domínio	Processamento Paralelo
Grupo	7
Sujeito	Objeto
Condor	Paralelismo funcional
Dag	Algoritmo ibl
Openmp	Apis
Sistema solaris	Host
Nó de destino / Nó de origem	Checkpoint provisório
modelo mpmd	Modelo spmd
cluster	Chaveamento de contexto
servidor	Corba
ferramenta de shell	Cluster
desenvolvedor	Poisson

Domínio	Modelagem Estocástica
Grupo	7
Sujeito	Objeto
san	san
modelo_san	hmm
multicast	algoritmo_split
soma_tensorial	produto_tensorial
módulo_pi	multicast
patching	controle_de_admissão
sgspn	patching
equilíbrio_de_nash	aunf
pepa	transmissão_multicast
árvore_de_atingibilidade	nacks

Domínio	Geologia
Grupo	8
Sujeito	Objeto
sedimentação	sedimentação
fácies	fácies
bacia	bacia
afloramento	eletrofácies
planície	lençol freático
sistema_fluvial	ângulo_de_contato
magmatismo	recristalização
metamorfismo	geminação
simulador_de_fluxo	espaço_poroso
lâmina	tensão_superficial

Domínio	Mineração de Dados
Grupo	8
Sujeito	Objeto
AUC	AUC
Rede som	Itemset
Apriori	Análise ROC
Spring	JADE
Ag	SMOTE
Lhs	r-tree
Especialista de domínio	Shapefile
DASE	Fuzzy
Algoritmo de clusterização	Análise de Pareto
Plataforma de agentes	Fonetograma

Domínio	Processamento Paralelo
Grupo	8
Sujeito	Objeto
cluster	cluster
mpi	mpi
escalonamento	escalabilidade
kernel	loop
pvm	speedup
openmp	aplicação paralela
daemon	ambiente paralelo
scheduler	corba
benchmarking	cardinalidade
condor	lei de amdahl

Domínio	Pediatria
Grupo	9
Sujeito	Objeto
bcg_id	tdah
lactente	disfunção_miccional
surfactante	mastite
noi	vhb
rnpt	dmo
aap	pic
enurese	incidência_de_dbp
anemia_falciforme	hic
aleitamento_materno	cardiomegalia
imc	concentração_de_iga

Domínio	Geologia
Grupo	9
Sujeito	Objeto
fácies	sedimentação
bacia	fácies
litologia	bacia
metamorfismo	cimentação
biotita	soerguimento
cimentação	ETRP
muscovita	biotita
footwall	dolomita
hangingwall	moscovita
soerguimento	eletrofácies