UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SOM4SIMD: UM MÉTODO SEMÂNTICO BASEADO EM ONTOLOGIA PARA DETECTAR SIMILARIDADE ENTRE DOCUMENTOS

CLAUDINEIA GONÇALVES DE ARRUDA

ORIENTADORA: PROFª. DRª. MARILDE TEREZINHA PRADO SANTOS

São Carlos - SP
Fevereiro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

CENTRO DE CIÊNCIAS EXATAS E DE TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

SOM4SIMD: UM MÉTODO SEMÂNTICO BASEADO EM ONTOLOGIA PARA DETECTAR SIMILARIDADE ENTRE DOCUMENTOS

CLAUDINEIA GONÇALVES DE ARRUDA

Dissertação/Qualificação/Tese apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação, área de concentração: Banco de Dados

Orientadora: Drª. Marilde Terezinha Prado dos Santos

São Carlos - SP

Fevereiro/2017

UNIVERSIDADE FEDERAL DE SÃO CARLOS

Centro de Ciências Exatas e de Tecnologia
Programa de Pós-Graduação em Ciência da Computação

Folha de Aprovação

Assinaturas dos membros da comissão examinadora que avaliou e aprovou a Defesa de Dissertação de Mestrado do candidato Claudineia Gonçalves de Arruda, realizada em 13/02/2017:

Profa. Dra. Marilde Terezinha Prado Santos
UFSCar

Profa. Dra. Marcela Xavier Ribeiro
UFSCar

Profa. Dra. Elis Cristina Montoro Hernandes
IFSP - São Carlos

Dedico essa dissertação a toda minha família e ao meu esposo.

AGRADECIMENTO

Primeiramente, agradeço a Deus por proporcionar-me essa oportunidade única.

Agradeço imensamente a minha querida orientadora Profª. Drª. Marilde Terezinha Prado dos Santos, por selecionar-me para ser sua orientanda, mesmo não me conhecendo. Agradeço Marilde, exatamente por tudo: todas as nossas reuniões, conversas e pela extrema paciência que teve em orientar-me e ensinar-me conteúdos que nunca pensei que poderia aprender. Além de orientadora, você foi e é uma amiga muito especial, dando-me ânimo nos momentos difíceis, mostrando-me que eu poderia ir além das minhas expectativas. Marilde, saiba que este título não é só meu, mas seu também, dedico-lhe de coração. Ser-lhe-ei eternamente grata.

Agradeço ao grupo de Educação Especial da UFSCar por ceder os documentos e ontologia que permitiram desenvolver este trabalho. Também agradeço pela disponibilidade em fazer os experimentos do mesmo.

Agradeço toda minha família (mãe, irmãos e avó), por todas as orações, mensagens de carinho e conforto, fazendo-me mais forte a cada dia. Sem vocês eu não estaria onde estou hoje, muito obrigada.

Agradeço muito ao meu amado marido Fernando, por toda força, incentivo e paciência que sempre teve comigo. Pelos conselhos adoráveis e por acreditar em mim mais do que eu mesma. Este título é nosso.

Agradeço a todos os meus amigos e colegas do Departamento de Computação, pelas conversas e desabafos, sempre me fazendo mais forte e confortando-me. Agradeço em especial ao Arthur, companheiro de todos os dias, que me ajudou muito em vários aspectos, até os últimos segundos. Agradeço muito à Mirela, que me recebeu muito bem e me ajudou bastante na chegada em São Carlos. Agradeço a minha amiga de anos, Mara, por ouvir meus desabafos e sempre me entender.

Agradeço a todos os professores do Departamento de Computação pelo conhecimento passado a mim durante as aulas, reuniões ou outros tipos de conversas. Saibam que sou muito grata por ter tido a oportunidade de conviver e

aprender um pouco com vocês. Eu nunca imaginei que poderia conhecer pessoas tão experientes e cheias de conhecimento. Serei eternamente grata a todos vocês.

Agradeço também a todos os demais funcionários do Departamento, pois o trabalho de vocês permitiu que eu pudesse concluir minha pesquisa.

RESUMO

Em diversas áreas de pesquisas, as entrevistas são um meio de obtenção de dados muito utilizadas por pesquisadores. Essas entrevistas são dispostas, na maioria das vezes, em diversos documentos e têm uma linguagem informal, por se tratar de conversas entre várias pessoas ao mesmo tempo. Analisar tais documentos é uma tarefa árdua e demorada, trazendo cansaço e dificuldades para uma análise correta. Uma solução para análise desse tipo de entrevistas é agrupar os documentos de acordo com a similaridade que existem entre eles, pois assim os especialistas conseguem analisar os documentos de assuntos parecidos de forma mais rápida. Desta forma, este trabalho apresenta o método SOM4SImD, criado para detectar a similaridade semântica entre os documentos compostos por entrevistas com uma linguagem informal escritas no português brasileiro. Para criar este método, foi utilizado uma ontologia de mesmo domínio dos documentos, que permitiu o uso dos termos formais da ontologia, juntamente com seus sinônimos e variantes para realizar a anotação semântica nos documentos e para realizar o cálculo da similaridade entre os pares de entrevistas. Através do método criado, foi desenvolvida uma abordagem SimIGroup que auxilia os pesquisadores na análise qualitativa dos documentos, utilizando a técnica *Coding*. Os resultados mostram que o método SOM4SImD e a abordagem SimIGroup diminuem as dificuldades e cansaço na análise dos documentos realizadas pelos anotadores, auxiliando no aumento da quantidade de documentos analisados. Além disso, o método SOM4SImD se mostrou mais vantajoso na obtenção de similaridade entre documentos do que os demais encontrados na literatura, alcançando valores significantes para as medidas de desempenho, com 0,96 de precisão, 0,93 de revocação e 0,94 de F-Mensure.

Palavras-chave: método, similaridade semântica, documentos, ontologia, abordagem, análise de documentos, análise qualitativa.

ABSTRACT

In several research areas, interviews are a means of obtaining data widely used by researchers. These interviews are arranged, in most cases, in several documents and have an informal language, because they are conversations between several people at the same time. Analyzing such documents is an arduous and time-consuming task, bringing fatigue and difficulties to a correct analysis. One solution for analyzing this type of interview is to group documents according to the similarity between them, so that experts can analyze documents of similar subjects more quickly. In this way, this work presents the method SOM4SImD, created to detect the semantic similarity between the documents composed by interviews with an informal language written in Brazilian Portuguese. In order to create this method, an ontology of the same document domain was used, which allowed the use of the formal terms of the ontology, along with its synonyms and variants, to perform the semantic annotation in the documents and to calculate the similarity between the interview pairs. Through the created method, a SimIGroup approach was developed that assists the researchers in the qualitative analysis of the documents, using Coding technique. The results show that the SOM4SImD method and the SimIGroup approach reduce the difficulties and fatigue in the analysis of the documents made by the annotators, helping to increase the number of documents analyzed. In addition, the SOM4SImD method was more advantageous in obtaining similarity between documents than the others found in the literature, reaching significant values for the performance measures, with 0.96 accuracy, 0.93 of recall and 0.94 of F-Mensure.

Keywords: method, semantic similarity, documents, ontology, approach, document analysis, qualitative analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

Capítulo 1

INTRODUÇÃO

1.1 Contexto e Motivação

Nos meios de pesquisas, em geral, há grandes quantidades de dados gerados, advindos de vários tipos ou formas, muitas vezes sendo por meio de entrevistas, questionários e outros gêneros textuais. Tais dados são armazenados em diversos formatos: arquivos de texto, áudio, vídeo e outros. As pesquisas desse tipo são realizadas por diversas áreas, desde as exatas até as humanas. O grande problema desses documentos é analisá-los, pois existe uma grande quantidade dos mesmos nos arquivos, e que normalmente, são falas livres, gerando um texto com palavras fora do padrão da língua portuguesa. Dependendo da análise e da área, torna-se muito cansativo, pois algumas vezes, é preciso fazer análise qualitativa dos dados, o que demanda muita atenção na leitura dos dados armazenados.

Existem várias instituições de ensino que têm informações armazenadas em suas bases de dados para a realização de pesquisas, uma delas é a Universidade Federal de São Carlos (UFSCar), que tem um curso de graduação/pós-graduação na área de educação especial. Este curso faz parte da área de humanas e contém vários grupos de pesquisas. Vários pesquisadores da educação especial da UFSCar fazem parte, juntamente com pesquisadores de outras instituições de ensino, de um projeto em nível nacional chamado Observatório Nacional da Educação Especial (ONEESP). Esses pesquisadores são de universidades localizadas em diversas regiões do Brasil. Os dados do ONEESP foram organizados em um banco de dados central, localizado na UFSCar e é composto por dissertações, teses, artigos,

entrevistas de professores que trabalham com educação especial, entre outros tipos de documentos.

Essa grande base de dados necessita de ser totalmente analisada, e o melhor instrumento analítico para esse tipo de dado, escolhido pelos pesquisadores, é a análise qualitativa. O processo envolvido na realização de análise qualitativa é um grande desafio para os pesquisadores, pois demanda muito tempo de várias pessoas que possuem conhecimento específico no tema.

Analisar esses documentos qualitativamente é difícil e demorado, demandando de vários especialistas de domínio. Neste âmbito, foi desenvolvido um método que encontra documentos similares, levando em consideração os termos presentes em uma ontologia de domínio. Dessa forma, sabe-se quais documentos têm conteúdos mais similares entre si. Além disso, este método mostra para os especialistas de domínio os termos mais importantes do domínio do texto, facilitando a análise qualitativa do mesmo.

Diante do contexto e motivação mencionados anteriormente, o objetivo geral desta pesquisa é empregar tecnologias computacionais, tais como anotação semântica, ontologia e processamento de língua natural para obter similaridade semântica entre documentos com grande volume de texto.

Esse trabalho apoia-se nas seguintes hipóteses:

1. É possível obter similaridade semântica entre documentos com grande capacidade de texto utilizando ontologia;

2. É possível auxiliar especialistas de domínio, no processo de análise qualitativa, identificando documentos com textos similares;

3. A obtenção de similaridade semântica entre documentos diminui o tempo de análise qualitativa do documento;

1.2 Metodologia de Desenvolvimento do Trabalho

O desenvolvimento desse trabalho de pesquisa está pautado nas seguintes atividades:

- Levantamento bibliográfico, utilizando o método de revisão sistemática, a fim de conhecer o assunto e verificar os trabalhos similares já desenvolvidos na área;
- Estudo e instanciação da abordagem escolhida para a realização da anotação semântica dos documentos;
- Definição e implementação dos algoritmos para desenvolvimento do método SOM4SIMD;
- Desenvolvimento de um protótipo que implemente o método proposto neste trabalho, incorporando o algoritmo de visualização semântica dos documentos.
- Definição e realização da experimentação com o protótipo desenvolvido, utilizando os dados da educação especial, armazenados na base de dados da ONEESP na UFSCar, visando o comparativo com análises manuais dos documentos;

A presente pesquisa se trata de um estudo de caso, onde é desenvolvido um método e após é feito um estudo de caso do próprio.

1.3 Organização do Trabalho

Esta dissertação está organizada da seguinte maneira: o Capítulo 2 é constituído da fundamentação teórica, ou seja, são explicados todos os pontos importantes utilizados no desenvolvimento deste trabalho. O Capítulo 3 contém os trabalhos correlatos a esta pesquisa, importantes para mostrar as diferenças e vantagens deste trabalho. O Capítulo 4 é dedicado ao método proposto, explicando o passo a passo utilizado para o desenvolvimento do método. O Capítulo 5 mostra os experimentos e resultados dos mesmos. Mostrando que o método auxilia em análises de documentos, explicando as limitações e vantagens do trabalho

desenvolvido. O Capítulo 6 é dedicado a conclusão e sugestões de trabalhos futuros que podem ser desenvolvidos a partir desta pesquisa.

Capítulo 2

FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações Iniciais

Este capítulo tem como propósito apresentar os assuntos abordados ao longo do desenvolvimento deste trabalho. A apresentação destes assuntos é de extrema importância para que o leitor entenda os mecanismos empregados na criação da pesquisa.

2.2 Ontologia

As ontologias têm sido empregadas para a representação do conhecimento, principalmente com o desenvolvimento da Web Semântica (BERNERS-LEE et al., 2001), que investiga como tratar a semântica do conteúdo e de serviços disponíveis na Web. Por representarem informação semântica e permitirem inferência de conhecimento, ontologias têm sido utilizadas em diversas aplicações para facilitar a comunicação: tanto entre humanos, quanto entre os sistemas computacionais (YAGUINUMA; SANTOS; BIAJIZ, 2007).

Pela definição de Gruber (1993), Borst (1997) e Studer, Benjamins e Fensel (1998): “Uma ontologia é uma especificação formal e explícita de uma conceitualização compartilhada”. Na Ciência da Computação, uma ontologia define

um conjunto de primitivas representacionais com as quais é possível modelar um domínio de conhecimento ou discurso (GRUBER, 2009). De modo geral, tais primitivas são hierarquias de conceitos (taxonomias), atributos, relacionamentos, indivíduos e restrições (axiomas), que descrevem o conhecimento de um determinado domínio de modo consensual e compartilhado (GUARINO, 1998).

De acordo com Noy e Mcguinness (2001), uma ontologia é uma descrição formal de conceitos ou classes em um domínio de discurso. As classes podem estar organizadas na forma de uma taxonomia, permitindo subclasses e superclasses. As propriedades descrevem as características e os atributos de cada conceito, podem ser divididas em propriedades de objetos (Object Property) e propriedades de tipos dados (DataType Property). A primeira determina relações binárias entre indivíduos ou classes; a segunda relaciona um indivíduo a um dado literal. As restrições sobre os conceitos e propriedades expressam significado de forma que máquinas sejam capazes de interpretar por meio de raciocínio automático. Por fim, os indivíduos ou instâncias de classes, junto à ontologia, constituem a base de conhecimento.

2.3 Anotação Semântica

O processo de vincular modelos semânticos e linguagem natural é conhecido como anotação semântica. Esse processo cria inter-relações entre ontologias e documentos não estruturados ou semiestruturados (LI; BONTCHEVA, 2007). A anotação semântica é a atribuição de links para a descrição semântica de cada entidade localizada nos documentos (KIRYAKOV et al., 2004).

Segundo Uren et al. (2006) as ferramentas que permitem a produção de anotações semânticas podem ser manuais ou automáticas. A anotação manual é feita por ferramentas que fornecem um editor de texto e suporte para ontologias, podendo possuir recursos que apoiam a anotação semântica, como é o caso da GATE. A anotação semiautomática pode ser feita por ferramentas que proporcionam sugestões para as anotações, porém, que ainda necessitam da intervenção de especialistas. As anotações automáticas podem ser feitas por ferramentas que realizam as anotações automaticamente. Geralmente esses processos são apoiados em técnicas de extração de informação baseadas em ontologias.

2.4 Processo de Língua Natural

Nessa dissertação foram utilizadas duas técnicas de processamento de língua natural: a Sentenciação e a Tokenização.

2.4.1. Sentenciação

A Sentenciação é o processo que segmenta o documento em sentenças. A sentença pode ser identificada quando é detectado um caractere de pontuação que marca o fim de uma sentença. Cada sentença é anotada com a marcação Sentence. Cada ponto final é marcado com uma marcação Split. (PEREIRA, 2014)

2.4.2. Tokenização

A tokenização é o processo que segmenta o documento em partes denominadas de tokens. Os tokens podem ser palavras, números, pontuação e outros, desde que seja uma unidade semântica útil para processamento. Uma palavra é uma sequência contínua de caracteres alfanuméricos, separada por espaços em branco ou por caracteres de pontuação. Da mesma forma, os números são definidos como uma sequência consecutiva de dígitos. Tanto a palavra, quanto os números são marcados com uma marcação Token que os identifica do primeiro ao último caractere. Os caracteres de pontuação são reconhecidos individualmente e marcados com a marcação Token. Os espaços em branco são marcados com a marcação SpaceToken.

Para Manning e Schütze(1999), um passo inicial de processamento é dividir o texto de entrada em unidades chamadas fichas, onde cada um é uma palavra ou outra coisa, como um número ou um sinal de pontuação. Este processo é referido como geração de tokens.

2.5. Similaridade Semântica

A similaridade Semântica mede é uma medida utilizada para verificar a semelhança entre pares de palavras, sentenças ou documentos. Este conceito de similaridade nasceu há muito tempo, quando se iniciaram as pesquisas sobre a semântica em frases e textos. Para falar em similaridade semântica, normalmente utiliza-se uma base de dados semântica, sendo uma ontologia ou qualquer outro tipo de base de dados do mesmo gênero.

Para este trabalho foram utilizadas duas medidas de similaridades semânticas desenvolvidas nos anos 90, quando as primeiras abordagens de similaridades semânticas surgiram. A medida de similaridade utilizada é a medida de Lin (1998). Essa medida é considerada híbrida porque mistura os conceitos de abordagens mais antigos.

A medida de similaridade utilizada para desenvolver o método proposto utiliza o conceito de Conteúdo da Informação. A abordagem baseada em Conteúdo da Informação indica a quantidade de informação fornecida por um conceito em um contexto (WU; PALMER, 1994). O conteúdo da informação de um conceito em um determinado corpus é definido na Equação 1.

$$IC(c) = - \log p(c) \quad (1)$$

O p(c) é a probabilidade de ocorrência do conceito c, ou qualquer ocorrência de conceitos que são generalizados por c, no corpus do domínio de estudo. Por exemplo: na taxonomia, o conceito “aluno” é especializado em “criança” e “adolescente”, desta forma a probabilidade leva em conta tanto o conceito “aluno” como os conceitos “criança” e “adolescente”. A Equação 2 mostra como é possível encontrar a probabilidade de um conceito no corpus.

$$P(c) = Freq(c)/N \quad (2)$$

Na Equação 2 a Freq(c) é a soma das ocorrências de c e seus conceitos especializados no corpus. O N é o total de conceitos presentes no corpus. Desta maneira, p(c) é o resultado soma das frequências de (c) dividido pelo número total

de conceitos no corpus. Quanto maior é a probabilidade de um conceito c ocorrer no corpus, ele se torna mais abstrato (um conceito mais geral tem a sua probabilidade acrescida de todos os seus mais específicos). E consequentemente, quanto mais abstrato, menor é o seu conteúdo de informação.

Outro fator importante é encontrar o conteúdo da informação do conceito que é o ancestral comum dos conceitos que estão sendo comparados, ou seja, encontrar o IC do conceito que é pai dos conceitos em comparação (SUSSNA, 1993). Normalmente, este conceito é representado pela sigla LCS. O LCS é encontrado na hierarquia da ontologia que está sendo utilizada para encontrar a similaridade semântica de conceitos. A Equação 3 mostra a fórmula de cálculo do conteúdo da informação do LCS.

$$Sim(c1,c2) = IC(LCS(c1,c2)) \quad (3)$$

Onde o Sim(c1,c2) é o resultado da similaridade semântica ente os conceitos c1 e c2. A abordagem de Lin (1998) propõe uma análise teórica do problema da similaridade, chegando a uma medida genérica que afirma ter uma abrangência a diferentes domínios que sigam um modelo probabilístico, tais como: valores ordinais, vectores de características, palavras, conceitos numa taxonomia, entre outros. A medida de similaridade de Lin (1998) tem como objetivo descobrir conceitos iguais em um corpus. A fórmula pode ser vista na Equação 4.

$$Sim\ (c1,c2) = 2*IC(LCS(c1,c2))/\ IC(c1) + IC(c2) \quad (4)$$

A medida de similaridade de Lin (1998) foi utilizada neste trabalho, pois ela se apresentou mais completa, por ser uma medida mais evoluída de outras existentes, e também por ser mais aplicável para a situação percebida nesta dissertação.

2.6. Análise Qualitativa

De acordo com Strauss e Corbin (2008) a pesquisa qualitativa busca o entendimento de um assunto especifico por meio de descrições, comparações e

interpretações dos dados, ao contrário da pesquisa quantitativa, que utiliza valores numéricos. Contudo, é importante ressaltar que é perfeitamente possível realizar análise qualitativa com valores numéricos.

Segundo Seaman (1999), uma das vantagens de utilizar métodos de análise qualitativa é que o pesquisador aprofunda-se na complexidade do tema ou problema estudado e não se preocupa com abstrações do problema. Esse tipo de análise exige maior esforço do pesquisador se comparado a métodos quantitativos.

De acordo com Hancock (2002) a pesquisa qualitativa é bastante empregada nas áreas humanas e médicas, por tratarem-se de áreas em que o comportamento humano é bastante avaliado. Seaman (1999) afirma que nos últimos anos, a pesquisa qualitativa também está sendo empregada como a tecnologia, já que o comportamento humano é muito avaliado em se tratando de analisar dados coletados onde o ser humano é altamente ativo.

Analisar corretamente os dados de uma pesquisa é uma tarefa considerada difícil; tanto para análise quantitativa, quanto para análise qualitativa. Quando se trata de analisar dados quantitativamente os tipos de coisas utilizadas são frequências de variáveis, testes estatísticos, probabilidades, entre outros. Quando se trata de analisar os dados qualitativamente, não utiliza os mesmos meios da análise quantitativa e sim outros métodos.

Com o dado textual em mãos, o pesquisador lê o texto em busca de referências sobre o tema de interesse da pesquisa e aplica um rótulo (código) para cada parte relevante do texto. O procedimento é o mesmo se os dados forem coletados de outras formas (HERNANDES, 2012). Esse método de análise é chamado de Análise de Conteúdo (*Contentanalysis*) (HANCOCK , 2002) ou *Coding* (SEAMAN, 1999).

Hernandes (2012) relata ainda que a Análise de Conteúdo ou *Coding,* envolve codificação e classificação de dados, em que a ideia básica é identificar, a partir dos dados coletados, trechos que são relevantes e que, de alguma forma, representam informações importantes escondidas em um conjunto de dados em forma de texto.

Para essa dissertação a análise qualitativa é um ponto que será beneficiado com o método criado. Pretende-se que o método criado auxilie os pesquisadores do estudo de caso na análise qualitativa de dados.

2.7. Medidas de Desempenho

As medidas de desempenho do método criado foram as métricas de Precisão, Revocação e F-Measure, utilizadas no campo de Recuperação de informação (FRAKES; BAEZA-YATES, 1992), que também podem ser utilizadas na Extração de Informação para medição de desempenho (COWIE; LEHNERT, 1996). E na Extração de Informação Baseada em Ontologia (MAYNARD; PETERS; LI, 2006).

Neste trabalho elas foram utilizadas nos experimentos para medir as anotações feitas pelos anotadores sem o auxílio do método criado e as anotações com o auxílio do método criado.

De acordo com Wimalasuriya e Dou (2010b), as medidas de desempenho podem ser representadas pelas seguintes fórmulas:

$$\text{Precisão} = \text{Anotações corretas} / \text{Total de Anotações} \quad (5)$$

$$\text{Revocação} = \text{Anotações corretas} / \text{Total de anotações no Gold Standard} \quad (6)$$

Precisão, Equação 5, mede o número de ocorrências corretamente anotadas como uma porcentagem do número de ocorrências anotadas, ou seja, mede quantas das ocorrências que o anotador ou a aplicação identificaram são realmente corretas, independentemente de não terem conseguido recuperar todos as ocorrências corretas. Quanto maior a precisão, melhor é a anotação e assegura que o que foi anotado é correto (PEREIRA, 2014).

Revocação, na Equação 6, mede o número de ocorrências corretamente anotadas como uma porcentagem do número total de ocorrências no Gold Standard, ou seja, mede quantas das ocorrências que deveriam ter sido anotadas foram realmente anotadas, independentemente de quantas anotações incorretas foram feitas. Quanto maior a taxa de revocação, melhor o desempenho do anotador ou da aplicação nas anotações de ocorrências corretas (PEREIRA, 2014).

2.8. Considerações Finais

Neste capítulo foram apresentados os assuntos que serão abordados ao longo do desenvolvimento do método desta dissertação. O uso de ontologias é de extrema importância neste trabalho, pois a utilização de uma ontologia de domínio permite que o método seja eficiente e inovador. Através da ontologia, é possível fazer a anotação semântica, que é de extrema importância, pois é assim que os documentos serão anotados em suas palavras e expressões importantes.

O uso de técnicas de Processamento de Língua Natural é importante, pois permite o pré-processamento dos documentos, resultando num documento que pode ser analisado.

A similaridade semântica é imprescindível neste trabalho, pois é um dos pontos principais do método criado, sendo utilizados métodos já existentes na literatura.

A análise qualitativa é citada neste capítulo, porque uma das contribuições do método criado é na área da análise qualitativa. As medidas de desempenho são utilizadas para medir o desempenho do método criado.

Capítulo 3

TRABALHOS CORRELATOS

.

3.1 Considerações Iniciais

Nessa seção serão apresentados os trabalhos com maior grau de similaridade com a presente proposta. Estes trabalhos foram selecionados em uma pesquisa feita nas bibliotecas de buscas através do método de revisão sistemática.

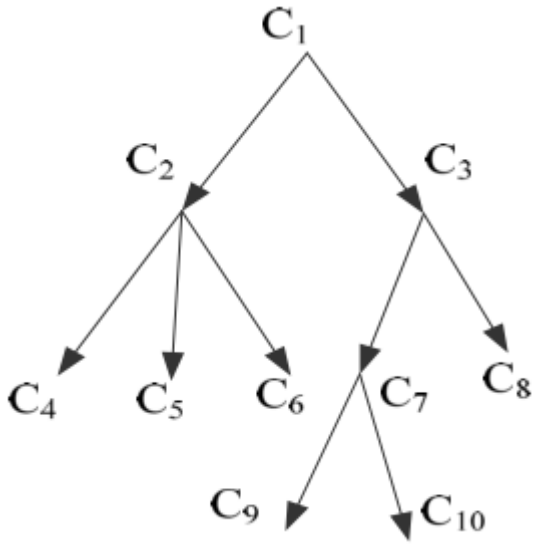
3.2 Avaliando Similaridade de Texto Usando Ontologia (LIU, WANG; 2014)

Liu e Wang (2014) desenvolveram um método de similaridade semântica entre sentenças e textos grandes. O método desenvolvido é primeiramente focado em similaridade entre sentenças e é aplicado também em textos, utilizando a distribuição de pesos para termos mais frequentes no texto. O método criado pelos autores utiliza uma ontologia, para identificar termos importantes presentes no texto.

Para desenvolver o método para encontrar a similaridade entre sentenças, os autores tomam por referência cinco definições. A primeira definição é a representação da ontologia em forma de uma árvore, facilitando a identificação das relações entre as classes ontológicas, sendo essa árvore a base do método proposto. A Árvore de Conceito Hierárquica (HCT, do inglês Hierarchical Concept

Tree) é denotada como T(N,E), onde N é o conjunto de nós da árvore e E é o conjunto de arestas entre os pares pais/filhos de N. A cobertura semântica do nó filho é a partição da cobertura semântica do seu nó pai. A HCT é um tipo de árvore baseada em ontologia. Na Figura 1 é possível ver a ontologia baseada em árvore.

Figura 1: Modelo de Ontologia baseada em Árvore

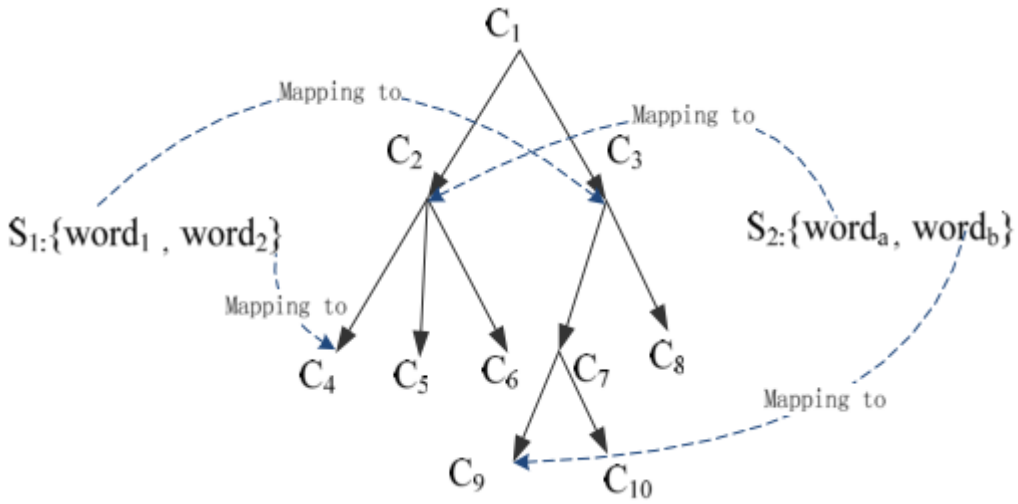


Fonte: Liu e Wang (2014)

A segunda definição determina os conceitos nós e suas relevâncias. Os nós de relevância de um nó conceito no HCT são o seu nó ancestral e nós descendentes. Nas sentenças existem muitas palavras com irrelevância semântica, como os artigos, preposições, entre outras, essas palavras são excluídas da sentença e as palavras que permanecem são consideradas palavras chaves das frases sendo mapeadas para a ontologia de domínio.

A terceira definição trata do conceito de sentença nós de ligação direta. Se uma palavra chave na sentença é equivalente a um nó na HCT, então o nó conceito é nomeado de nó de ligação direta. Os nós de ligação direta de todas as expressões que compõe as sentenças são chamados de sentença nós de ligação direta. Na Figura 2 pode ser visto o mapeamento dos nós de ligação direta na ontologia.

Figura 2: Mapeamento de ligações de nós



Fonte: Liu e Wang (2014)

A quarta definição é sobre sentenças nós relevância. Os nós de relevância de uma determinada frase no HCT são a união de todos os nós de relevância de cada nó de ligação direta da frase.

Por fim, a quinta definição refere-se ao conceito de nós de ligação direta e nós relevância com base em vetores conceito. Dado um HCT com N conceitos, o vetor conceito de uma frase é denotado como S= (v1, v2, ..., vn) e Vi= (i = 1, 2, ... n) é o valor da dimensão correspondendo a todos os nós do conceito da ontologia em relação ao par de frases. A fórmula para cálculo do vetor de similaridade pode ser vista na Equação 7.

$$sim(V\_1, V\_2) = V\_1 * V\_2 / ||V\_1|||||V\_2|| \quad (7)$$

Onde V_1 e V_2 são os vetores 1 e 2, respectivamente.

A similaridade entre documentos é definida por um conjunto de documentos nós de ligação e documentos nós de relevância. O cálculo da similaridade dentre os documentos utiliza a medida de ponderação tf–idf. Essa é uma medida estatística usada para avaliar o quão importante uma palavra é em um documento. A importância aumenta proporcionalmente ao número de vezes que uma palavra ocorre no documento, mas é compensada pela frequência da palavra no corpus. Duas intuições estão em jogo no sistema de ponderação tf-idf: quanto mais uma palavra aparece no texto, mais importante ela é; em quanto mais documentos a

palavra ocorre, menos discriminador é este documento, ou seja, menor é sua contribuição na caracterização da semântica de um documento.

O peso das palavras nos documentos foi extraído da seguinte forma: Para cada taxonomia no vetor conceito, o algoritmo recursivamente propaga pesos para o nó pai até que o nó de raiz é alcançado. A fórmula, extraída do artigo é dada na Equação 8.

$$W_{pai} = a_2 * W_{filho} \quad (8)$$

Onde W_{pai} é o peso do pai e W_{filho} é peso do filho e a_2 é o fator de propagação de peso.

O fator a_2 propagação de peso é usado para determinar o quanto de peso de uma palavra filha é propagada para seu pai.

O método desenvolvido pelos autores consegue encontrar sentenças e documentos similares utilizando a similaridade entre os termos da ontologia pela hierarquia horizontal e vertical. O trabalho não utiliza os relacionamentos presentes na ontologia para definir similaridade entre sentenças. No final do processo, é usado o vetor de cosseno de similaridade e não faz uso da ontologia.

3.3 Similaridade de Sentenças Baseada em Vetor de Modelo Semântico (JINGLING, HUIYUN; 2014)

Jingling, Huiyun e Baojiang (2014) desenvolveram um trabalho que encontra a similaridade semântica entre sentenças, utilizando vetor de palavras. O método proposto pelos autores é aplicável a textos pequenos, levando em consideração informações semânticas, estruturais e de ordem das palavras em uma sentença.

O método proposto pelos autores é dividido em três etapas: primeiramente é obtida a similaridade semântica das palavras. No segundo passo é obtida a similaridade semântica entre a sentença baseada na similaridade das palavras e a estrutura das sentenças. Por fim, é calculada a similaridade de ordem das palavras entre as sentenças e combinada com a similaridade entre as sentenças. Para o cálculo de similaridade entre as palavras é utilizado um dicionário de corpus.

A similaridade entre palavras é realizada através de uma base de conhecimentos léxicais chamado How-net. O How-net define uma palavra como uma linguagem de descrição de um conhecimento multidimensional. Cada palavra é como uma unidade muito pequena que tem uma relação de cima para baixo de modo que todas as palavras formem uma árvore de hierarquia. A distância entre as palavras na taxonomia é utilizada para calcular a similaridade entre elas.

A similaridade semântica entre as sentenças é realizada utilizando um vetor de palavras, sendo tais palavras presentes nas frases. Dados duas sentenças, são formados dois vetores de palavras, o vetor 1 e o vetor 2. Para estes vetores são feitos vários tipos de tarefas a fim de definir as palavras que serão realmente utilizadas na comparação final. Ao final do processo resulta-se em um vetor semântico, onde a similaridade entre duas frases semântica pode ser definida como o coeficiente de cosseno entre os dois vectores semânticos.

Para definir a similaridade de uma sentença, além da semântica entre as palavras, também é levado em consideração a ordem em que as palavras estão dispostas na sentença. Nesta fase são identificadas as palavras que são verbos e pronomes e trabalhadas em vetores para realizar a similaridade do cosseno. A fórmula para definir o coeficiente de cosseno entre dois vetores é mostrada na Equação 9, onde S_1 significa a palavra 1 e S_2 significa a palavra 2.

$$Sim\_p = (S\_1 * S\_2) / (|S\_1| * |S\_2|) \quad (9)$$

A ordem em que as palavras estão dispostas nas frases interfere no significado das mesmas. No cálculo da similaridade de ordem das palavras nas sentenças, as palavras são analisadas e mapeadas de acordo com sua posição na sentença. Para medir essa semelhança, os autores propõem a medida que pode ser vista na Equação 10.

$$Sim\_g = 1- (||r\_1 - r\_2||) / (||r\_1 + r\_2||) \quad (10)$$

Por fim é calculada a semelhança entre sentenças, levando em consideração a semelhança entre as palavras e a semelhança entre a ordem das palavras. Desta forma, a medida de similaridade mostrada na Equação 11, mostra como calcular a similaridade entre as sentenças.

$$Sim = e * Sim\_p + (1 - e) * Sim\_g \quad (11)$$

3.4 Medida de Similaridade Usando a Abordagem de Conteúdo de Informação com Profundidade para o Cálculo da Similaridade (GUPTA, YADAV; 2014)

No trabalho de Gupta e Yadav (2014) é proposta a criação de uma métrica para o cálculo da relação semântica entre pares de conceitos, utilizando a abordagem baseada em características de Tversky, que leva em consideração a característica comum e distinta dos dois termos ou conceitos. Se a semelhança é maior comparada com as diferenças, a similaridade entre os conceitos é alta, caso contrário a similaridade é baixa. A teoria de Tversky é quantificada pelo conteúdo de informação de dois conceitos e o conteúdo de informação do ancestral comum mais específico de dois conceitos. À medida que é percorrido a hierarquia do WordNet, há um conceito mais específico e mais informativo. Enquanto mais é percorrida a altura da hierarquia, encontram-se mais Conceitos Generalizados e menos Informativos. Portanto, a profundidade de um conceito na hierarquia WordNet é um fator crítico no cálculo da similaridade. Levou-se em consideração a profundidade do conceito específico na hierarquia WordNet que é o fator decisivo para determinar a relevância de característica distinta específica de um conceito no cálculo de similaridade.

Para desenvolver a métrica proposta, os autores utilizaram o conteúdo de informações da Resnik (1995) e as abordagens baseadas em Tversky para o cálculo da similaridade. Foram introduzidos novos parâmetros α e β que levam em consideração a relevância de característica distinta específica para cada conceito com base na profundidade. A métrica proposta é mostrada nas Equações 12, 13 e 14.

$$Sim(new) = 3.Ic(lcs(c1,c2)- \alpha\ IC(c1)- \beta.IC(c2)) \text{ quando } ci \neq c2 \quad (12)$$

$$\alpha = depth(c1) / depth(c1)+depth(c2) \quad (13)$$

$$\beta = depth(c2) / depth(c1)+depth(c2) \quad (14)$$

Nas equações apresentadas, c1 e c2 são os termos que estão sendo comparados. IC é conteúdo da informação dos termos e depth é a profundidade de cada termo.

Os autores explicam que, se ambos os conceitos estão no mesmo nível que o valor de $\alpha = \beta$, o que implica que característica distinta de ambos os conceitos são igualmente significativos para o cálculo da similaridade. Se a diferença entre a profundidade dos conceitos é alta, $\alpha > \beta$ ou $\beta > \alpha$, então o conceito que está em maior profundidade será mais significativo para contribuir nas diferenças de similaridade. Na medida de similaridade criada, foi introduzida a relevância de características distintas, específica ao conceito, com base em sua profundidade na hierarquia. Com a introdução do parâmetro α e β, a medida de similaridade semântica proposta mostra melhora no termo de precisão, especialmente para pares de palavras com maior grau de similaridade. À medida que foi introduzida a relevância de uma característica distinta ao conceito com base na sua profundidade na taxonomia, o antepassado comum mais específico entre dois conceitos que representa a semelhança entre dois conceitos, encontra a ponderação sobre a característica distinta no caso de valores com alto grau de semelhança.

Os autores fizeram experimentos e mostraram que, em comparação com medidas já consolidadas na literatura, as medidas criadas por eles são mais eficientes dos que as de autores renomados na área, mostrando grau de acurácia maior, em relação à outras medidas existentes.

No trabalho apresentado nessa seção, os autores utilizam a base de dados semântica WordNet, muito extensa, abrangendo diversos domínios, sendo utilizados por diversos outros trabalhos. Os autores não utilizam grande quantidade de texto, como documentos, para encontrar a similaridade e sim textos menores e mais formais.

3.5. Considerações Finais

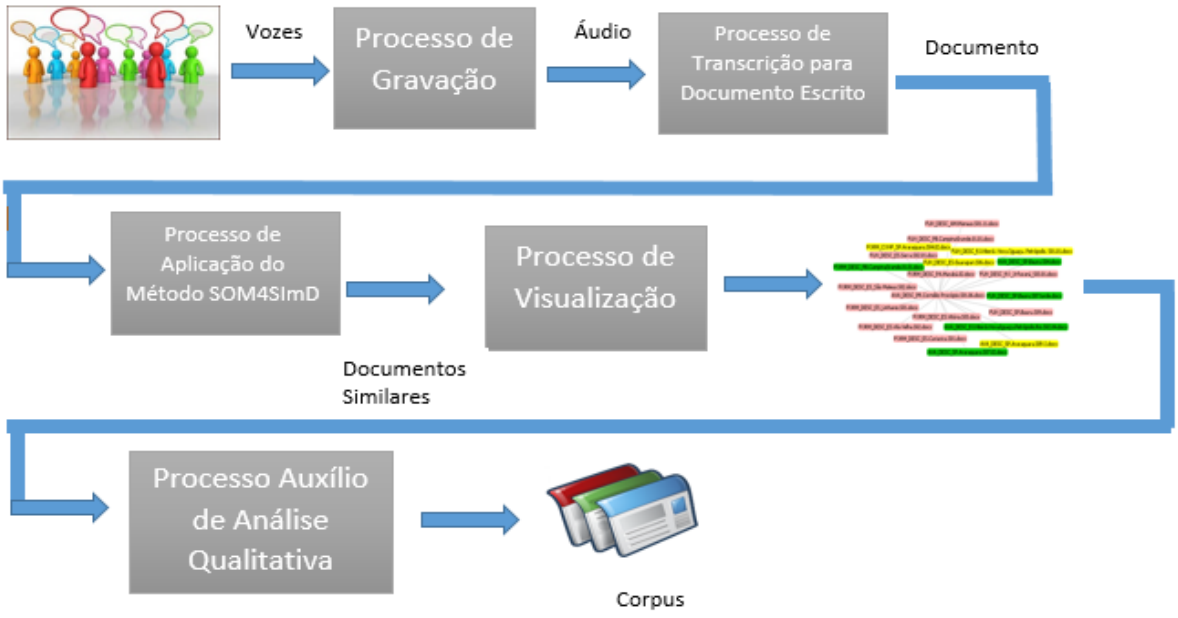
Neste capítulo foram apresentados os trabalhos que são correlatos ao trabalho desenvolvido nesta dissertação. Todos os trabalhos mostrados têm como objetivo obter similaridade semântica entre texto, utilizando ontologia. Contudo, nenhum deles faz uso dos termos da ontologia, considerando os sinônimos e variantes, apenas considera a hierarquia da ontologia, diferentemente do método proposto nesta pesquisa.

Capítulo 4

SIMIGROUP - ABORDAGEM DE AGRUPAMENTO DE ENTREVISTAS SIMILARES

A abordagem Similar Interviews Group approach (SimIGroup), Abordagem para agrupamento de entrevistas similares, contempla a aplicação do método SOM4SImD e outros processos desenvolvidos por pesquisadores de diversas áreas. A abordagem é mostrada na Figura 3.

Figura 3: Abordagem SimIGroup



Fonte: Autor

O ponto inicial da abordagem, mostrada na Figura 3, refere-se à reuniões que ocorrem entre pesquisadores de diversas áreas, que muitas vezes são chamadas de grupo focal. Nessas reuniões são discutidos os temas de interesse do grupo focal, podendo ser feitas através de perguntas e respostas. Normalmente, em discussões de grupos focais as entrevistas ou conversas são feitas de forma informal.

As conversas entre as pessoas presentes na reunião são gravadas através de um aparelho que grava apenas áudio, guardando assim as vozes das pessoas que dissertam durante as reuniões.

Após o processo de gravação, tem-se o áudio gravado das reuniões, que passam por outro processo, que transcreve o áudio para documentos de texto. Durante essa transcrição, a escrita ocorre de acordo com a fala das pessoas, sendo assim escritas, algumas vezes, fora dos padrões formais da língua portuguesa falada no Brasil.

Findado o processo de transcrição de áudio, têm-se vários documentos de texto compostos pelas entrevistas ocorridas nas reuniões. Esses documentos são compostos por textos de tamanhos variados, dependendo das conversas ocorridas nas reuniões. Tais documentos são fornecidos como entrada para o processo de aplicação do método SOM4SImD, que se constitui como contribuição científica desse trabalho para a área da computação. Este método é explicado detalhadamente na próxima seção.

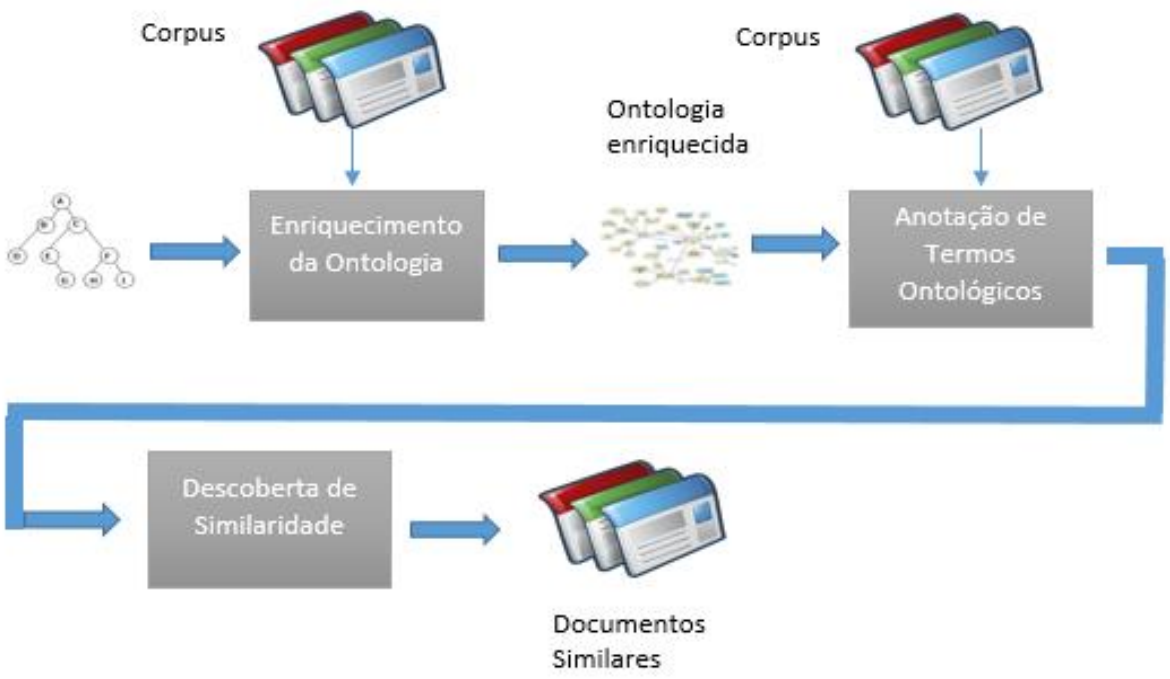
A saída do processo de aplicação do método SOM4SImD são documentos com algum grau de similaridade entre si. Esses documentos são fornecidos como entrada para o processo de Visualização, que neste trabalho foi instanciado com a utilização de um algoritmo disponibilizado na ferramenta *Prefuse* (http://www. http://prefuse.org/). Essa ferramenta gera a rede de visualização dos documentos exemplificada na Figura 3.

4.1 Método SOM4SImD

O método *SEMANTIC ONTOLOGY-BASED METHOD FOR DETECTING SIMILARITIES AMONG MULTIPLE DOCUMENTS* (SOM4SimD), é composto pelos

processos Anotação de termos Ontológicos e Descoberta de Similaridade, conforme apresentado na Figura 4. Para a implementação do método foi utilizada a linguagem de programação Java e algumas técnicas de Processamento de Língua Natural para facilitar o processamento dos documentos.

Figura 4: Método SOM4SImD



Fonte: Autor

O ponto inicial do método é a utilização de um corpus composto por vários documentos com as transcrições do áudio das entrevistas. Este corpus passa por um processo de anotação semântica baseada em ontologia, que foi desenvolvido baseado no trabalho de Pereira (2014). Este corpus anotado é fornecido como entrada para o processo de descoberta de similaridade semântica, que agrupa os documentos de acordo com as similaridades entre eles.

Corpus

O corpus de estudo do domínio é constituído de entrevistas realizadas com profissionais da área de educação especial de escolas de todo o Brasil. Estas

entrevistas estão em formato digital, são descrições de áudio e estão no idioma Português do Brasil.

Na Figura 5 é mostrado um extrato de um dos documentos do corpus utilizado. Nesta Figura percebe-se a informalidade da língua portuguesa que foi utilizada nas entrevistas. A informalidade dos documentos deve-se ao fato dessas entrevistas serem transcritas de áudios feitos de reuniões de grupos focais, onde o diálogo entre os pesquisadores foi feito de forma informal. As entrevistas foram feitas entre profissionais da educação especial em várias regiões do Brasil, aumentando a informalidade dos documentos devido ao regionalismo presente na fala. Ao todo, são 270 documentos que fazem parte do corpus estudado. Todos estes documentos estão divididos em três tipos de assuntos dentro do domínio: Avaliação, Formação e Funcionamento.

Figura 5: Extrato de um documento do corpus

Sílvia:
É para facilitar também o censo na hora de preencher o censo.

Colaboradora 2 de grupo focal 2

Em relação aos tipos de serviços, há vários tipos. Hoje o município tem atividade com futebol, não sei se vocês já viram a quadra que também tem natação, tem atendimento com a fono na SEMED, tem o atendimento com a terapeuta ocupacional, que é o básico não é? Apenas esses três serviços básicos que tem atendimento. Neuro|psicológico não tem. Avaliação, orientação tem, mas atendimento lá no local não tem.

Coordenadora de grupo focal 2

É agendado ou são os pais que levam para o atendimento na SEMED?

Colaboradora do Grupo focal 2:

Com a fono? É preciso os pais esperarem muito.

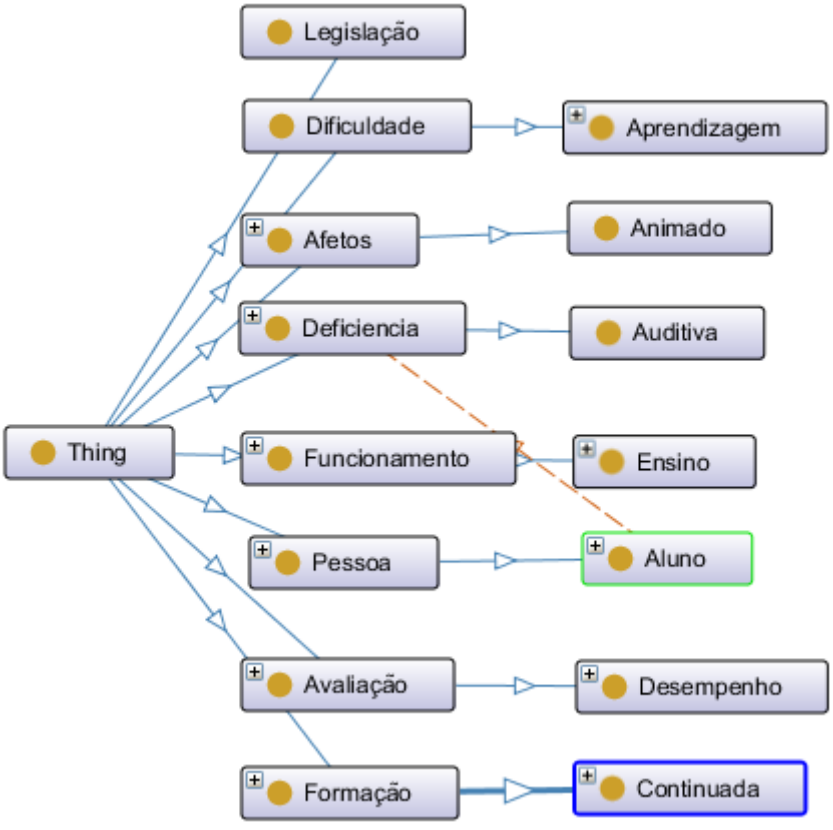
Coordenadora de grupo focal 2
Em alguns lugares, não sendo o caso de Marabá, tem transportes para estar levando essas crianças para garantir que esses alunos tenham acesso a esse tipo de serviço.

Fonte: Extraído do Corpus

Ontologia

Para o desenvolvimento deste trabalho foi utilizada a ontologia Educação Especial, desenvolvida como parte da tese de doutorado de Fernandes (2015) para o domínio do contexto da educação especial no Brasil. O contexto da educação especial no Brasil engloba as políticas vigentes, pessoas participantes, ensino, deficiência dos alunos e recursos utilizados, entre outros. Na Figura 6 é mostrado um extrato da ontologia, evidenciando, apenas suas classes mais gerais.

Figura 6: Extrato da Ontologia Educação Especial



Fonte: Fernandes (2015)

A classe Legislação refere-se à legislação vigente em torno da educação especial no Brasil, essa é única classe do nível 1 da taxonomia da ontologia que não é especializada. A classe Deficiência refere-se aos tipos de deficiência que os alunos podem possuir. A classe Funcionários refere-se aos tipos de funcionários que trabalham com o domínio estudado. A classe Avaliação refere-se aos tipos de

avaliação de ensino impostas aos alunos. A classe Dificuldade refere-se aos tipos de dificuldades referentes à aprendizagem dos alunos. A classe Afetos refere-se ao tipo de sentimento, negativo ou positivo, relacionado ao aluno. A classe Pessoa, refere-se a todas as pessoas envolvidas no contexto. A classe formação refere-se à formação dos profissionais.

Essa ontologia é composta por 94 conceitos, que são as classes da ontologia. A classe *Thing*, que também está inserida nessa contagem, é a classe mais geral, de nível zero, que é a raiz de todas as outras classes.

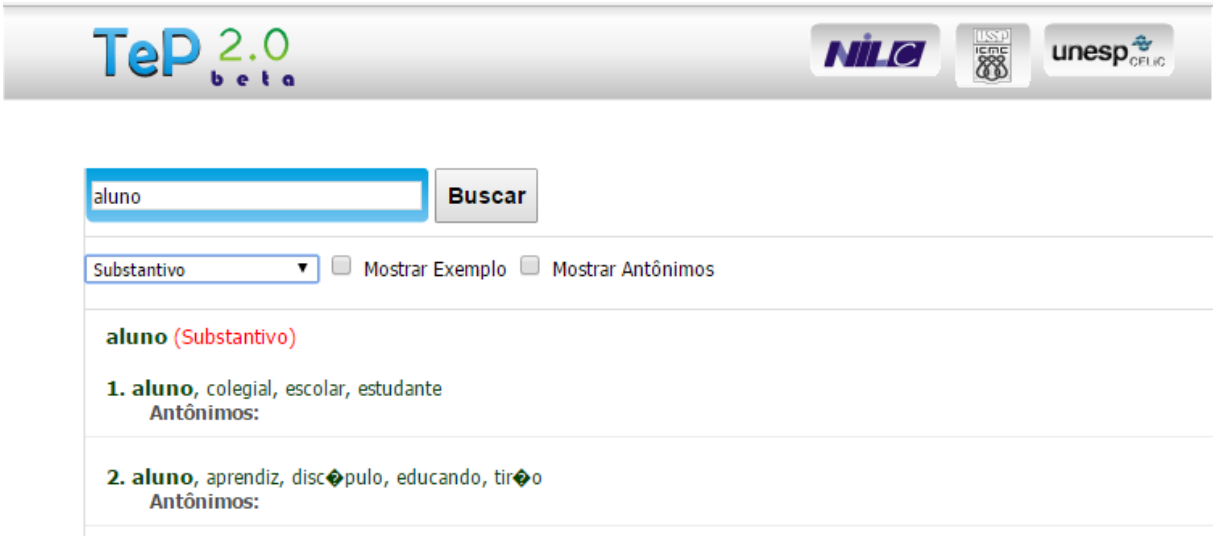
Enriquecimento da Ontologia

Todas as 94 classes presentes na ontologia são classes de grande importância no domínio da educação especial, contudo apenas as classes na ontologia não abrangem todas as formas de dialeto presentes nos documentos de estudo de caso dessa dissertação. Dessa forma a ontologia foi enriquecida com sinônimos e variações de cada classe da ontologia.

Para o enriquecimento da ontologia, foi feito, inicialmente, uma reunião com os especialistas de domínio para obter os termos similares de cada classe da ontologia. Essa reunião foi essencial, pois os especialistas de domínio conhecem bem os documentos e puderam fornecer os termos similares usados nas entrevistas.

Após a reunião, também foi consultado um tesauro da língua portuguesa do Brasil para obter sinônimos das classes da ontologia. O tesauro utilizado foi o TeP 2.0 Beta, desenvolvido por membros do grupo de pesquisa Nilc (http://www.nilc.icmc.usp.br/tep2/) da Universidade de São Paulo (USP). Um exemplo de pesquisa realizada no Tep é mostrado na Figura 7.

Figura 7: Exemplo de pesquisa no Tep



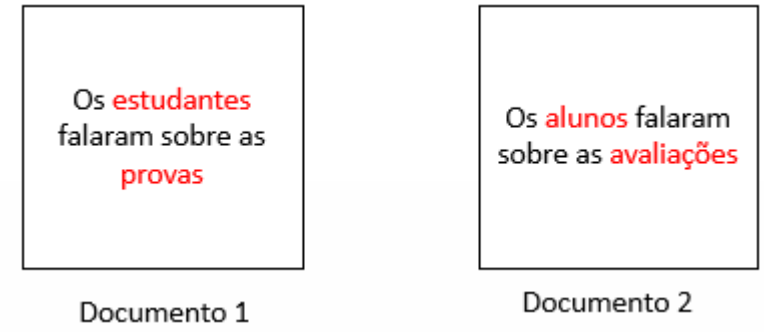
Fonte: Nilc (2015)

Após a obtenção dos sinônimos fornecidos pelos especialistas de domínio e pelas consultas no Tep, todos os sinônimos e variantes das classes ontológicas foram inseridos na ontologia Educação Especial como instâncias.

Os métodos para detectar similaridade entre documentos encontrados na literatura consideram apenas a hierarquia vertical e horizontal da ontologia, e não suas instâncias.

Com o uso das instâncias, considerando os termos sinônimos e variações dos termos ontológicos, considera-se que é possível obter uma similaridade mais precisa do que utilizando somente termos ontológicos e suas comparações. A Figura 8 mostra uma ilustração dessa comparação.

Figura 8: Comparação de Documentos



Fonte: Autor

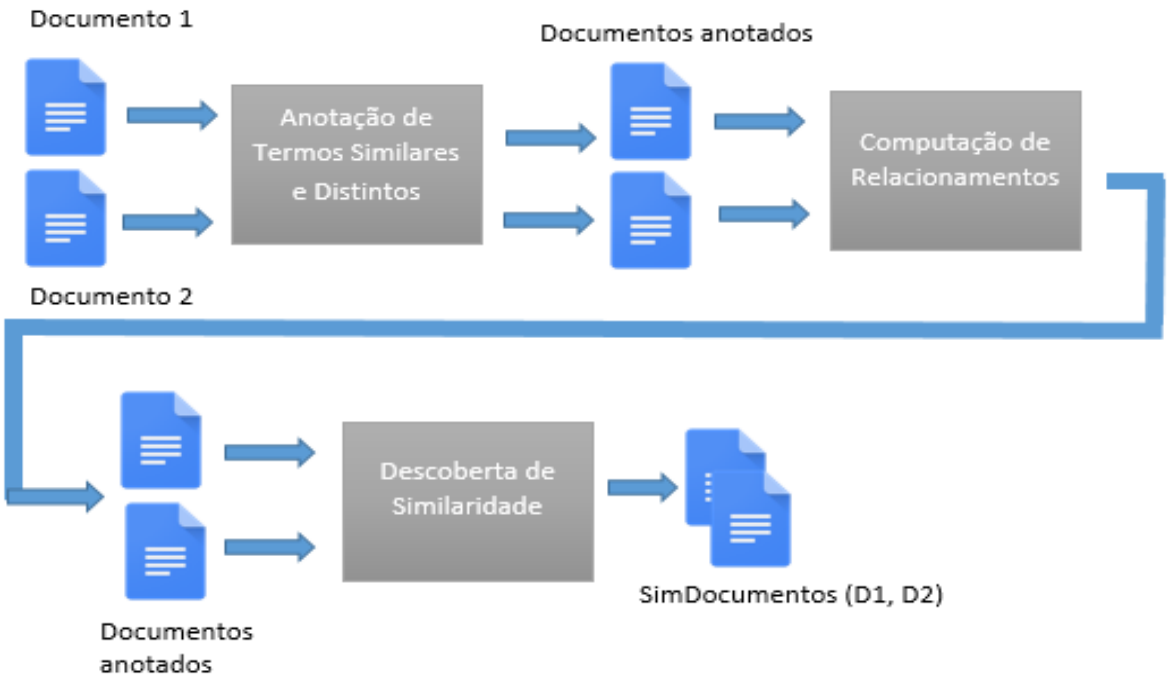
Pelos métodos tradicionais, os documentos mostrados na Figura 8, contêm uma similaridade pequena. Já no método proposto, os documentos possuem uma

similaridade bem maior, se comparados com os métodos tradicionais. Isso devido ao fato do método proposto considerar estudantes e alunos, avaliações e provas, como palavras idênticas.

Anotação dos Termos Ontológicos

O processo de anotação dos termos ontológicos nos documentos, mostrando se estes termos são idênticos ou diferentes, pode ser visto na Figura 9. Nesta Figura também são vistos o processo de Computação de Relacionamentos e o processo de Descoberta de Similaridade.

Figura 9: Anotação de Termos Ontológicos

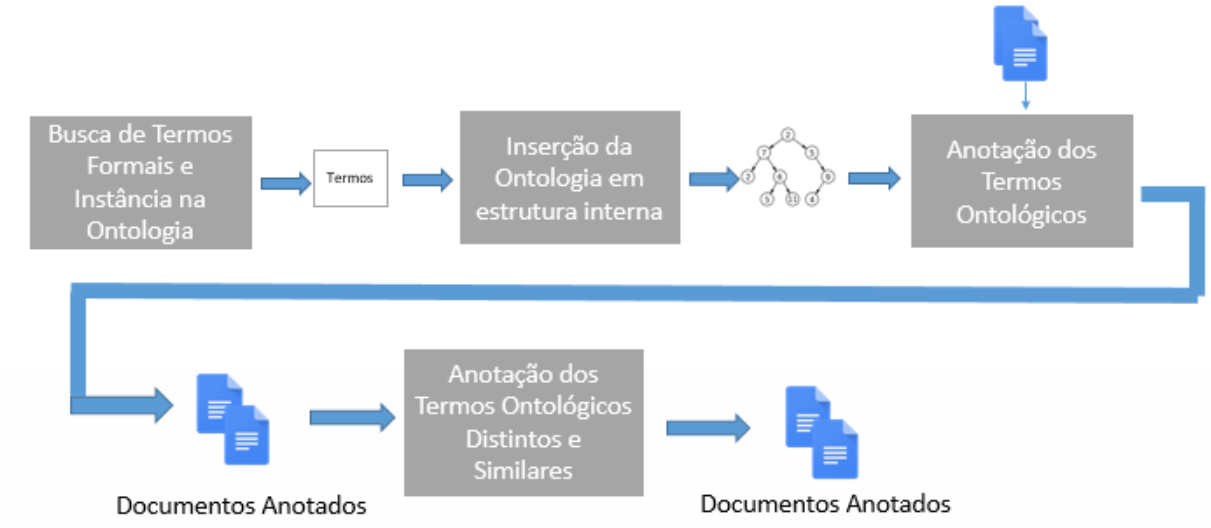


Fonte: Autor

Anotação dos Termos Similares e Distintos

O Processo de anotação dos termos similares e distintos é mostrado na Figura 10.

Figura 10: Processo de Anotação de Termos Similares e Distintos



Fonte: Autor

Para identificar os termos ontológicos nos documentos, seus sinônimos (similares) e variações, inicialmente deve ser obter a lista de termos oriundos da ontologia. Em nosso estudo de caso, essa lista foi obtida a partir de uma consulta de termos realizada sobre a Ontologia Educação Especial utilizando a ferramenta *Protegè* e a linguagem *SparQL Query*. Um exemplo de consulta pode ser visto na Figura 11.

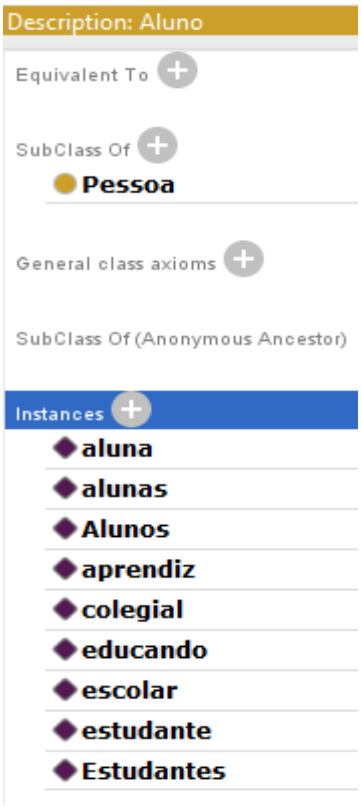
Figura 11: Exemplo de consulta na ontologia Educação Especial com um extrato do resultado

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
SELECT ?subject
                WHERE { ?subject rdfs:subClassOf ?object }
```

Diagnóstico
didáticos
Gestão
Processo
Comum
Adolescente
Triagem
tecnológicos
adaptação_de_materiais
Habilitação
Médico
Identificação

Fonte: Protegé

Figura 12: Instâncias da Classe Aluno



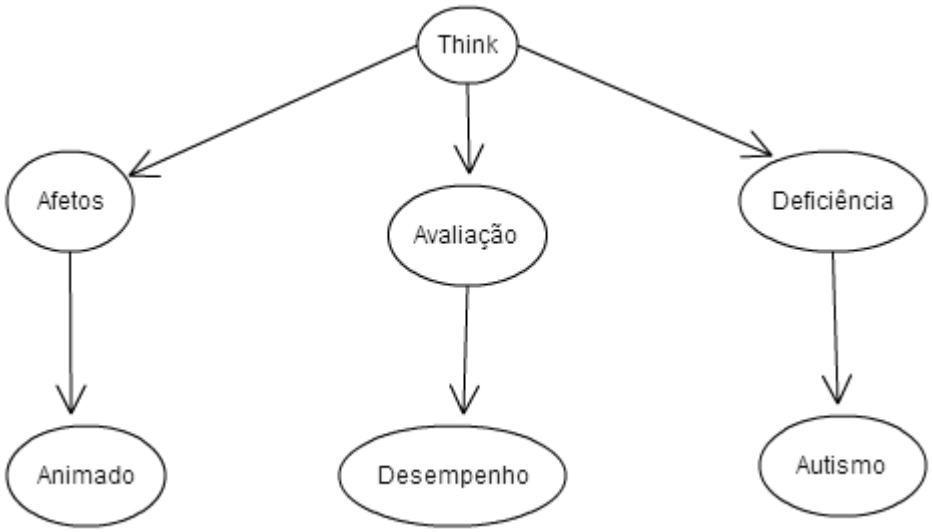
Fonte: Protegé

Na lista mostrada na Figura 11, apresentam-se apenas classes presentes na ontologia, contudo é preciso obter os sinônimos e variantes de cada termo. Essa obtenção também feita na Protegé é mostrada na Figura 12.

Após a obtenção da lista de termos ontológicos, seus variantes e sinônimos, esses termos foram implementados internamente utilizando a estrutura de dados árvore, onde as classes da ontologia nesta árvore foram chamadas de vértices e os relacionamentos de hierarquia da ontologia foram chamados arestas. O uso dessa estrutura de dados facilitou a representação da relação de pai/filhos presentes na ontologia. Todos os vértices têm uma lista de sinônimos e variantes que serão utilizados em passos mais avançados do método. A Figura 13 mostra uma ilustração de um extrato da ontologia na implementação.

As classes da ontologia são mostradas em formas de vértices. Uma classe pode ser pai e filha ao mesmo tempo, excluindo dessa afirmação a classe Thing, que é classe raiz e apenas classe pai. Na implementação foram criadas as classes Vértice e Aresta. As Figuras 14 e 15 mostram a implementação das classes Vertice e Aresta na linguagem Java.

Figura 13: Extrato da ontologia em forma de árvore



Fonte: Autor

A classe Vertice tem um atributo chamado nome, que dará nome ao vértice, uma lista de arestas e uma lista dos termos sinônimos e variantes, chamada de Similar. Os métodos addAdj adiciona uma aresta na árvore. A classe Aresta é composta por dois atributos do tipo Vertice, um com o nome de origem e outro com o nome de destino, formando as arestas através da Origem e do Destino. Desta forma é possível que a ontologia seja representada através das relações entre as classes Vértice e Aresta.

Os termos variantes e similares aos termos da ontologia também foram colocados na estrutura de dados interna. Um exemplo disso são os termos “aluno”, “alunos”, “estudantes” considerados neste trabalho como sinônimos entre si. Dessa forma, as variantes e sinônimos dos termos foram inseridas na ontologia e após consulta na mesma, obteve-se os termos sinônimos e variantes que foram implementados em uma classe chamada Similar.

Através da ontologia representada em forma de vértices e arestas, foi possível realizar o processor de anotação dos termos ontológicos nos documentos.

O primeiro passo do processo Anotação de Termos Ontológicos é o processo Anotação de Termos Similares e Distintos, conforme mostrado na Figura 9. Nesse processo as técnicas de Sentenciação e Tokenização são utilizadas nos documentos do corpus para prepará-los para a fase de identificação de termos ontológicos nos documentos.

Figura 14: Classe Vértice

```
public class Vertice {
    String nome;
    List<Aresta> adj;
    List<Similar> similares;

    Vertice(String nome) {
        this.nome = nome;
        this.adj = new ArrayList<Aresta>();
        this.similares = new ArrayList<Similar>();
    }

    void addAdj(Aresta e) {
        adj.add(e);
    }

    public String getNome() {
        return nome;
    }

    public void setNome(String nome) {
        this.nome = nome;
    }
}
```

Fonte: Autor

Figura 15: Classe Aresta

```
public class Aresta {
    Vertice origem;
    Vertice destino;

    Aresta(Vertice origem, Vertice destino) {
        this.origem = origem;
        this.destino = destino;

    }

    public Vertice getOrigem() {
        return origem;
    }

    public void setOrigem(Vertice origem) {
        this.origem = origem;
    }

    public Vertice getDestino() {
        return destino;
    }

    public void setDestino(Vertice destino) {
        this.destino = destino;
    }
}
```

Fonte: Autor

Os documentos são processados, utilizando as técnicas de Sentenciação e Tokenização implementadas na biblioteca *OpenNLP*. A Sentenciação é utilizada para dividir o documento em sentenças para permitir uma melhor aplicação da Tokenização, que foi utilizada para a comparação de cada palavra na sentença com os vértices do grafo e aos termos similares (sinônimos e variantes) ao do vértice, a fim de identificar quais são iguais aos vértices ou seus similares, portanto, que se referem aos conceitos ontológicos.

Os documentos processados foram marcados com *tags* para que pudessem ser analisados posteriormente. Foi desenvolvido um documento XML *Schema* para explicar os significados e comportamentos das *tags*. A Figura 16 mostra o extrato do documento XML *Schema Annotation*, com o comportamento das *tags* presentes nas marcações dos documentos anotados.

Figura 16: Extrato do documento XML *Schema Annotation*

```
<xsd:complexType name="REType">
<xsd:sequence>
 <xsd:element name="RS" type="RSType" minOccurs="0" maxOccurs="unbounded"/>
 <xsd:element name="RE" type="REType" minOccurs="0" maxOccurs="unbounded"/>
 <xsd:element name="TS" type="TSType" minOccurs="0" maxOccurs="unbounded"/>
 <xsd:element name="TD1" type="TD1Type" minOccurs="0" maxOccurs="unbounded"/>
 <xsd:element name="TD2" type="TD2Type" minOccurs="0" maxOccurs="unbounded"/>
 <xsd:element name="To" type="ToType" minOccurs="0" maxOccurs="unbounded"/>
</ xsd: sequence>
</ xsd: complexType>

<xsd:complexType name="RSType">
<xsd:sequence>
 <xsd:element name="To" type="ToType" minOccurs="0" maxOccurs="unbounded"/>
</ xsd: sequence>
</ xsd: complexType>

<xsd:complexType name="REType">
<xsd:sequence>
 <xsd:element name="To" type="ToType" minOccurs="0" maxOccurs="unbounded"/>
</ xsd: sequence>
</ xsd: complexType>

<xsd:complexType name="TSType">
<xsd:sequence>
 <xsd:element name="To" type="ToType" minOccurs="0" maxOccurs="unbounded"/>
</ xsd: sequence>
</ xsd: complexType>
```

Fonte: Autor

Na Figura 16 é mostrado que os documentos poderão ser anotados com até cinco tipos de *tags*: RS, RE, TS, TD1, TD2 e To. A *tag* RE pode ou não existir e se existir, ela conterá as *tags* TS, TD1, TD2 e To. A *tag* RE identifica os relacionamentos em um documento. As *Tags* TS, TD1 e TD2 são *tags* com mesmo nível e podem ou não existir. A *tag* TS significa Termo Similar, a *tag* TD1 significa que o termo é diferente e presente apenas no documento 1 e a *tag* TD2 significa que o termo é diferente e presente apenas no documento 2. Todas essas *tags*, TS, TD1 e TD2, se existirem, têm como *tag* filha a To. A *tag* To significa Termo Ontológico, e dentro dela tem um valor no formato de *string*, ou seja, palavras, que neste *Schema* foi definido um tamanho máximo de 30 caracteres.

Os termos ontológicos encontrados nos documentos são marcados nos próprios documentos com a *tag* <To> </To> mostrando que são palavras que se referem a termos ontológicos, como pode ser visto no exemplo de um extrato de documento anotado na Figura 17.

Figura 17: Extrato do documento com anotação de termos ontológicos

FORM_INFO_PR.Comélio Procópio. S01.02.xml

```
1  Referência do documento original: GF_PR_Cornélio
2  Procópio_2_12p_FORM.
3  Referência do documento modificado: FORM_DESC_PR.Cornélio
4  Procópio.S01.01
5  A política de inclusão escolar trouxe ou não novas demandas
6  para a <To>formação</To>  de <To>professores</To> de
7  <To>educação especial</To>? Se trouxe, quais? Política
8  de inclusão escolar, então nós vamos lembrar da
9  resolução 02 de 2001 e do decreto 17 e depois,
10 também da política de 2008. Os documentos que nós
11 temos mais específicos da inclusão escolar. Se ela trouxe novas
12 demandas pra <To>formação</To> dos <To>professores</To>,
13 nesse tema agora é <To>formação</To> de <To>professores</To>.
14 D – Não. Desde que a gente vem estudando as diretrizes,
15 desde que a gente vem vendo a política em si, nós não vimos
16 nenhuma capacitação de <To>formação</To>, de orientação melhor.
```

Fonte: Autor

Anotação de Termos Similares e Distintos

A anotação dos termos similares e distintos ocorre após a anotação dos termos ontológicos nos documentos. Para realizar a anotação dos termos similares e distintos é feita uma comparação entre pares de documentos, onde um documento é comparado com todos os demais do corpus.

Para anotar os termos ontológicos idênticos, foi utilizada a *tag* <TS> </TS>, e para anotar os termos ontológicos distintos foram utilizadas as *tags* <TD1> </TD1>, para termos ontológicos presentes no documento 1 e ausentes no documento 2; e <TD2> </TD2> para os termos ontológicos presentes no documento 2 e ausentes no documento 1. Exemplos de anotação de termos ontológicos similares e distintos são vistos nas Figuras 18 e 19.

Figura 18: Extrato do documento 1 com anotação de termos ontológicos similares e distintos

A segunda questão é: como você acha que deve ser a <TS><To>formação</To></TS> inicial do <TS><To>professor</To></TS> da <TD2><To>sala</To></TD2> de recurso multifuncional? Na cabeça de vocês, se vocês tivessem que pensar algo pra <TS><To>formação</To></TS> inicial, o que vocês acham que um <TS><To>professor</To></TS> deveria ter na sua <TS><To>formação</To></TS> inicial? Pra ser <TS><To>professor</To></TS> de <TD2><To>sala</To></TD2> de recurso multifuncional?
ProfaS: Eu acho que deveria ter um <TD2><To>curso</To></TD2> pra esses iniciantes.
Julia: você quer tentar especificar, detalhar um pouco melhor que tipo de <TD2><To>curso</To></TD2>, como que você acha que deve ser, tem alguma particularidade, só <TD2><To>curso</To></TD2>? ProfaS: Ele teve a <TS><To>formação</To></TS>, ai ele vai prestar o concurso, às vezes ele vai lá pra <TD2><To>sala</To></TD2> também sem aquela experiência, de como está trabalhando, então é ter uma outra <TS><To>formação</To></TS> pra esses <TS><To>professores</To></TS> que são os <TD2><To>cursos</To></TD2>.

Fonte: Autor

Figura 19: Extrato do documento 2 com anotação de termos ontológicos similares e distintos

Referência do documento original: GF_PR_Cornélio Procópio_2_12p_FORM.
Referência do documento modificado: FORM_DESC_PR.Cornélio Procópio.S01.01
A política de inclusão escolar trouxe ou não novas demandas para a <TS><To>formação</To></TS> de <TS><To>professores</To></TS> de <TD1><To>educação especial</To></TD1>?
Se trouxe, quais? Política de inclusão escolar, então nós vamos lembrar da resolução 02 de 2001 e do decreto 17 e depois, também da política de 2008. Os documentos que nós temos mais específicos da inclusão escolar. Se ela trouxe novas demandas pra <TS><To>formação</To></TS> dos <TS><To>professores</To></TS>, nesse tema agora é <TS><To>formação</To></TS> de <TS><To>professores</To></TS>.
D – Não. Desde que a gente vem estudando as diretrizes, desde que a gente vem vendo a política em si, nós não vimos nenhuma capacitação de <TS><To>formação</To></TS>, de orientação melhor.

Fonte: Autor

O algoritmo que realiza o processo de marcação dos termos ontológicos, termos similares e termos distintos entre os documentos é mostrado no Algoritmo 1.

Algoritmo 1: PROCESSO DE MARCAÇÃO DE TODOS OS TERMOS ONTOLÓGICOS ENTRE DOCUMENTOS

Entrada: *B1* e *B2* Documentos, A Estrutura Interna Ontologia
Saída: C Documentos anotados com termos ontológicos distintos e semelhantes
Declaração:

a_i iésimo termo presente em A
to_j jésimo termo ontológico presente no Documento B1
to_n enésimo termo ontológico presente no Documento B2

Início:

```
1   Sentencia B1
2   Sentencia B2
3   Tokeniza B1;
4   Tokeniza B2;
5   para cada a_i ϵ A faça
6       para cada to_j ϵ B_1 faça
7           Se a_i = to_j então
8               B1 ← tags(<To> </To>)
9               Lb1 ← Sentença
10              para cada to_n ϵ Lb1 faça
11                  V1[i].to[i] ← To
12                  V1[i].pont[i] ←to_n
13              fim
14          fim
15      fim
16      para cada to_n ϵ B_2 faça
17          Se a_i = to_n então
```

18 B2 ← tags(<To> </To>)
19 Lb2 ← Sentença
20 **para** cada to_n ϵ Lb2 **faça**
21 V2[i].to[i] ← To
22 V2[i].pont[i] ←to_n
23 **fim**
24 **fim**
25 **fim**
26 **fim**
27 Selection Sort (V1)
28 Selection Sort (V2)
29 **para** cada to_{token1} ϵ V1 **faça**
30 **para** cada to_{token2} ϵ V2 **faça**
31 **Se** to_{token1} = to_{token2} **então**
32 B1 ← <TS> </TS>
33 B2 ← <TS> </TS>
34 **fim**
35 **senão**
36 B1 ← <TD1> </TD1>
37 B2 ← <TD2> </TD2>
38 **fim**
39 **fim**
40 **fim**
41 **Retorna**: B1 e B2 anotados

A entrada do Algoritmo 1 é composta pelos dois documentos que serão comparados, e a estrutura de dados que contêm os termos ontológicos. Conforme as linhas 1 e 2 do algoritmo, esses documentos são sentenciados tokenizados e assim verificados se existem em seu texto termos ontológicos ou seus similares e variantes. Conforme linhas 3 a 18 do Algoritmo 1 ocorre a marcação de termos ontológicos nos documentos. Para todo termo ontológico presente no texto, este é marcado com as *tags* <To> </To> com a palavra entre as *tags*. Para marcar os termos ontológicos no texto, foram feitos laços para verificar a existência de termos ontológicos e seus similares no texto. Caso existam esses termos ontológicos no

texto, os documentos recebem as *tags* <To></To>. Após a marcação dos termos ontológicos nos documentos, a sentença onde está presente o *token* é colocada em uma lista, sendo cada *token* em uma posição da lista. Após, cada *token*, presente na lista, que representa o termo que representa os termos ontológicos é colocado em um vetor bidimensional que guarda o termo ontológico e seu endereço de *token*. Este processo ocorre para ambos os documentos, como vistos nas linhas 8 e 15 do algoritmo 1. Após a construção dos vetores de termos ontológicos e seus respectivos endereços, tais vetores são ordenados utilizando um método de ordenação. Para este trabalho foi utilizado o método *Insert Sort,* conforme linhas 19 e 20, mas poderia ter sido utilizado qualquer outro presente na literatura. Esses dois vetores de termos ontológicos são comparados para descobrir os termos iguais entre eles. Se existir os termos iguais entre os vetores, é feita a marcação com a *tag* <TS> </TS> nos dois documentos, conforme linhas 24 e 25. Essa marcação é possível, pois cada termo tem seu índice guardado no vetor. Caso os termos forem distintos são feitas as marcações que representam os termos distintos nos seus documentos de origem, utilizando a *tag* <TD1> </TD1> para marcar os termos distintos no documento 1 e a *tag* <TD2> </TD2> para marcar os termos distintos no documento 2, conforme linhas 28 e 29.

Computação de relacionamentos

O processo de computação de relacionamentos ocorre de maneira idêntica à marcação de termos similares e distintos. Para obter os relacionamentos presentes na ontologia, foi utilizada a ferramenta *Protegé*, que disponibiliza os relacionamentos. Exemplos de marcação desses relacionamentos são vistos na Figura 20.

Após as marcações de termos, os documentos anotados entram no processo de computação de relacionamentos. Através dos relacionamentos obtidos na consulta feita à ontologia, é possível buscar esses relacionamentos nos documentos.

Na Figura 20 as palavras que fazem parte do relacionamento ontológico são marcadas com as *tags* <RE> </RE>. No extrato do documento essas palavras são "professor", "ter" e "formação", formando o relacionamento "professor ter formação".

O algoritmo utilizado para buscar os relacionamentos na sentença dos documentos foi um algoritmo desenvolvido por Andrade (2017), em seu trabalho de mestrado, e adaptado para este trabalho, mostrado no Algoritmo 2.

Figura 20: Extrato de documento com marcação de relação ontológica

```
A segunda questão é: como você acha que deve ser a
<TS><To>formação</To></TS> inicial do
<TS><To>professor</To></TS> da <TD2><To>sala</To></TD2>
de recurso multifuncional? Na cabeça
de vocês, se vocês tivessem que pensar algo pra
<TS><To>formação</To></TS> inicial, o
que vocês acham que um <RE><TS><To>professor</To></TS></RE>
deveria <RE>ter</RE> na sua <RE><TS><To>formação</To></TS></RE> inicial?
Pra ser <TS><To>professor</To></TS> de
<TD2><To>sala</To></TD2> de recurso multifuncional?
ProfaS: Eu acho que deveria ter um <TD2><To>curso</To></TD2>
 pra esses iniciantes.
Julia: você quer tentar especificar, detalhar um pouco melhor
que tipo de <TD2><To>curso</To></TD2>, como que você
acha que deve ser, tem alguma particularidade,
só <TD2><To>curso</To></TD2>? ProfaS: Ele teve a
<TS><To>formação</To></TS>, ai ele vai prestar o
concurso, às vezes ele vai lá pra <TD2><To>sala</To></TD2>
também sem aquela experiência, de como está trabalhando,
então é ter uma outra <TS><To>formação</To></TS> pra esses
<TS><To>professores</To></TS> que são os
<TD2><To>cursos</To></TD2>.
```

Fonte: Autor

Após a obtenção dos relacionamentos ontológicos, é preciso anotá-los nos documentos. Para isso, também foi utilizado um algoritmo de Andrade (2017) com adaptação.

Algoritmo 2: PROCESSO DE ANOTAÇÃO DE RELACIONAMENTOS (Adaptado de Andrade, 2017)

Entrada: indexSn1, indexVerb, indexSn2, sentença

Saída: Documentos B com sentenças anotadas com as *tags* RE

1 **Início**
2 listaTokens ← tokeniza (sentenças)
3 **Para** token ϵ listaTokens **faça**
4 **Se** (indexSn1, indexVerb, indexSn2) contém index[i] **então**
5 newSent ← newSent + <RE> +token[i]+</RE>
6 V1[i].to[i] ← To

7 V1[i].pont[i] ←to_n
8 **senão**
9 newSent ← newSent + index[i]
10 **fim**
11 **Retorna** Documento anotado, Vetor de relacionamento

O algoritmo tem como entrada os index, que é o endereço de cada termo do relacionamento e a sentença referente ao relacionamento extraído. Essa sentença é novamente tokenizada. A lista de *tokens* é percorrida em busca pelos índices dos termos do relacionamento. Ao encontrá-los, o *token* é incorporado a *tag* <RE> </RE> e é concatenado na nova sentença. Caso contrário o termo é concatenado na nova sentença. Este processo também ocorre para o documento 2.

O algoritmo 3 mostra como ocorre a notação de relacionamentos similares nos documentos em comparação. A entrada são os dois documentos que estão sendo comparados e os dois vetores, de ambos os documentos de relacionamento. Após a construção dos vetores de termos ontológicos e seus respectivos endereços, tais vetores são ordenados, conforme linhas 2 e 3 do algoritmo 3, utilizando o método de ordenação *Insert Sort.* Esses dois vetores de termos ontológicos são comparados para descobrir os termos iguais entre eles. Se existir os termos iguais entre os vetores, é feita a marcação com a *tag* <RS> </RS> nos dois documentos, conforme linhas 24 e 25.

Algoritmo 3: PROCESSO DE MARCAÇÃO DE RELACIONAMENTOS ONTOLÓGICOS SIMILARES

Entrada B1 e B2 Documento, V1 e V2 vetores de relacionamento
Saída C Documentos anotados com relacionamentos semelhantes

1 **Inicio**
2 Selection Sort (V1)
3 Selection Sort (V2)
4 **para** cada to_{token1} ϵ V1 **faça**
5 **para** cada to_{token2} ϵ V2 **faça**
6 **Se** to_{token1} = to_{token2} **então**
7 B1 ← <RS> </RS>
8 B2 ← <RS> </RS>

```
9                   fim
10              fim
11      fim
12      Retorna B1 e B2 anotados
```

Descoberta de Similaridade

No processo de descoberta de similaridade é onde ocorrem os cálculos para descobrir o grau de similaridade entre os documentos. A entrada deste processo são os documentos anotados com os termos ontológicos similares e distintos entre os documentos e também os relacionamentos ontológicos.

Primeiramente é realizada a similaridade entre os termos distintos. Para o cálculo de similaridade entre os termos distintos é utilizada a medida de similaridade semântica de Lin (1998), explicada no capítulo 2. O processo de similaridade semântica entre os termos diferentes dos documentos só acontece se existirem termos diferentes entre os documentos. Para fazer este cálculo verifica-se nos vetores de termos distintos os termos presentes. Cada termo de um vetor é comparado com os demais do outro vetor, a fim de descobrir a similaridade entre eles. Após encontrar as similaridades entre todas as comparações, elas são somadas, resultando numa similaridade total entre os termos distintos dos documentos.

As saídas dos processos anteriores são entradas para a etapa de processo de cálculo da similaridade entre os documentos. Para distribuição de pesos é necessário fazer a contagem de termos iguais entre os documentos. Com esta contagem é feita a distribuição de pesos entre os termos. A similaridade entre documentos é a soma da similaridade entre os termos diferentes, a quantidade de termos idênticos e quantidade de relacionamentos idênticos entre os dois documentos. Desta forma é feita uma contagem de termos idênticos e relacionamentos idênticos. Se não houver termos diferentes entre os documentos, conclui-se que os documentos são idênticos, recebendo similaridade de 100%.

Algoritmo 4: PROCESSO DE OBTENÇÃO DE SIMILARIDADE ENTRE OS DOCUMENTOS 1 E 2

Entrada B1 e B2 Documentos V1 e V2 vetores de termos distintos

Saída SimAB Similaridade Semântica entre os documentos B1 e B2

Inicio

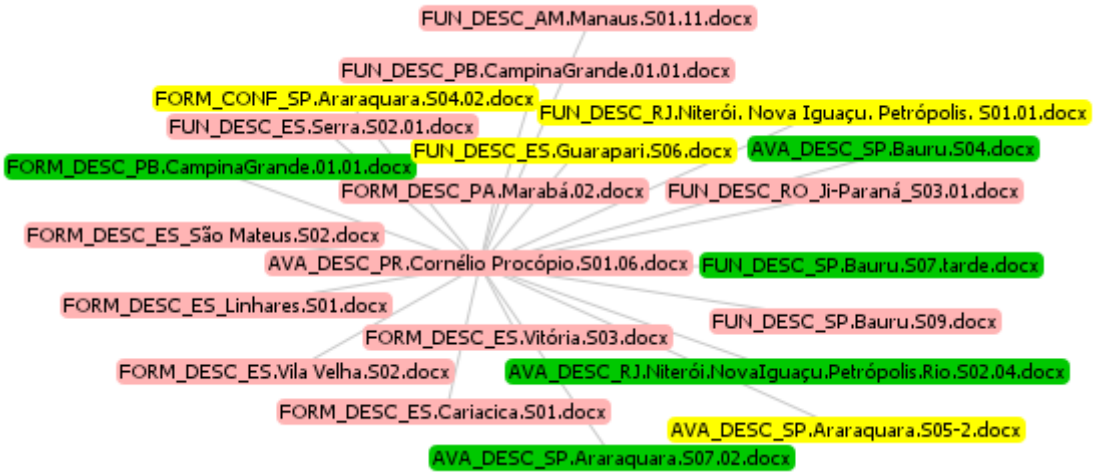
```
1   Se (V1 é V2 são vazios) então
2       SimAB ← 100;
3    fim
4    Senão
5       Para cada (<TS>  </TS> ϵ B1) faça
6           aux2 ←  aux2 +1;
7       fim
8       Para cada (<RS> </RS> ϵ B1) faça
9      aux3 ←  aux3 +1
10     fim
11     Se (aux2 < 5) então
12      SimAB ← SimTD + 0.3
13     fim
14     Se ((aux2 > 5) & (aux2 < 10)) então
15      SimAB ← SimTD + 0.5;
16     fim
17     Se (aux2 >= 10) então
18      SimAB ←  SimTD + 0.8;
19     fim
20     Se (aux3 > 0) então
21      SimAB ← SimA + 0.5;
22     fim
23    fim
24   fim
25  Retorna Sim AB
```

Processo de Visualização

O Processo de Visualização ocorre utilizando a Ferramenta Prefuse. A ferramenta Prefuse é um software extensível para ajudar na criação de aplicativos de visualização de informações interativas usando a linguagem de programação Java. Ele pode ser usado para criar aplicativos autônomos, componentes visuais incorporados em aplicativos maiores e *Applets Web*. Com a utilização desta ferramenta é possível criar gráficos, tabelas, árvores e várias outras estruturas para a representação de dados. O Prefuse foi desenvolvido em 2007, por grupos de pesquisadores de várias universidades dos Estados Unidos. Os dados completos são mostrados no site da ferramenta, no link http://prefuse.org/.

Após a obtenção da similaridade entre os documentos, esses valores são armazenados na estrutura de dados fila, onde são guardados os nomes dos documentos e seu valor de similaridade. Através da Lista montada com os documentos, foi utilizado um dos diversos algoritmos disponível na ferramenta prefuse, o *radialgraphview*, que através de nós disponíveis, faz a associação entre esses nós através de cores. Na atual pesquisa utilizou-se as cores verde, para par de documentos com similaridade acima de 85%, amarelo, para pares de documentos com similaridade entre 46% e 85% e rosa, para pares de documentos com similaridade de até 45%. Um exemplo da rede de documentos similares é visto na Figura 21.

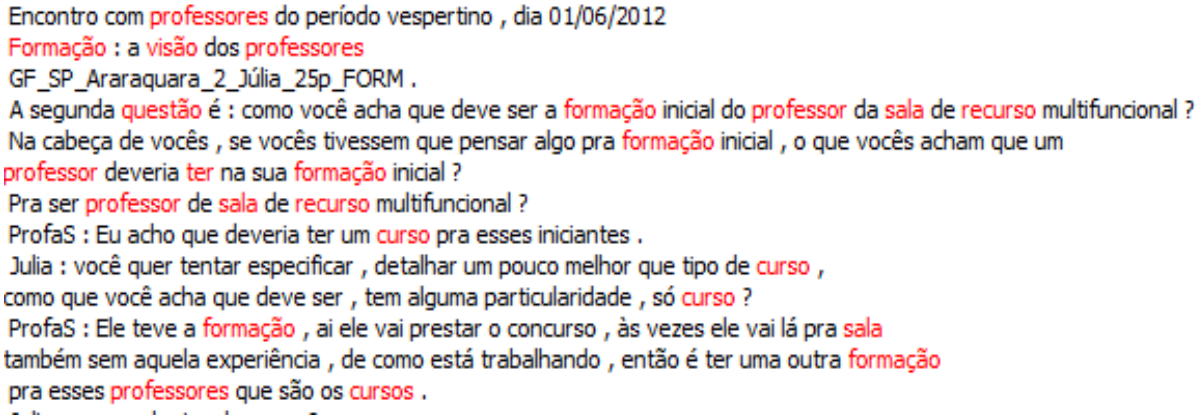
Figura 21: Rede de Documentos Similares



Fonte: Prefuse

A partir da rede semântica é possível clicar em cada documento que abrirá com os termos e relações ontológicas marcadas. Acredita-se que esta marcação feita nos documentos auxilia os pesquisadores a fazer análise qualitativa nos documentos, levando em consideração a técnica de *Coding*, onde marcam-se as palavras ou orações consideradas importantes nos documentos. Um exemplo de documentos com os termos principais marcados é mostrado na Figura 22.

Figura 22: Extrato de Documento marcado


Encontro com professores do período vespertino , dia 01/06/2012
Formação : a visão dos professores
GF_SP_Araraquara_2_Júlia_25p_FORM .
A segunda questão é : como você acha que deve ser a formação inicial do professor da sala de recurso multifuncional ?
Na cabeça de vocês , se vocês tivessem que pensar algo pra formação inicial , o que vocês acham que um professor deveria ter na sua formação inicial ?
Pra ser professor de sala de recurso multifuncional ?
ProfaS : Eu acho que deveria ter um curso pra esses iniciantes .
Julia : você quer tentar especificar , detalhar um pouco melhor que tipo de curso , como que você acha que deve ser , tem alguma particularidade , só curso ?
ProfaS : Ele teve a formação , ai ele vai prestar o concurso , às vezes ele vai lá pra sala também sem aquela experiência , de como está trabalhando , então é ter uma outra formação pra esses professores que são os cursos .

Fonte: Autor

Capítulo 5

EXPERIMENTOS E RESULTADOS

Para comprovar a relevância do método criado, este foi utilizado pelos especialistas de domínio da área da Educação Especial. Os experimentos foram realizados com três especialistas.

5.1. Protocolo de Experimentos

O protocolo de experimentos foi desenvolvido com o intuito de desenvolver um plano de experimentos simples para os especialistas de domínio e que tenha um resultado que comprove a eficiência e eficácia do método desenvolvido.

Os experimentos foram divididos em três etapas: a primeira, completamente manual, a segunda manual, mas com suporte dos termos da ontologia e a terceira, semiautomática, com o método já aplicado.

Primeira Etapa

A primeira etapa é feita totalmente manual, onde os especialistas de domínio têm um tempo pré-definido e analisam os conjuntos de documentos a fim de medir a

quantidade de documentos analisados em um determinado tempo. Para realizar a primeira etapa são necessários 3 especialistas e um corpus composto de aproximadamente 24 documentos.

Essa etapa é composta por dois processos: O primeiro processo é a criação do *gold standard* inicial, que é realizado por um especialista de domínio. Neste processo, os documentos serão analisados em pares, ou seja, o primeiro documento será comparado com todos os demais do corpus (exemplo: o primeiro documento com o segundo documento, o primeiro documento com o terceiro, e assim até terminar a lista de documentos). Para realizar este processo o especialista de domínio seguirá os seguintes passos:

- Ler o par de documentos escolhido com atenção, mas sem fazer nenhuma anotação;
- Ler novamente o par de documentos escolhido, anotando as palavras idênticas que encontrar nos dois documentos. Para a anotação de palavras idênticas, não são considerados os artigos, os numerais, as conjunções e as preposições.
- Ler pela terceira vez o par de documentos a fim de encontrar mais palavras idênticas em ambos;
- Anotar os números dos documentos e a quantidade de palavras idênticas encontradas;
- Realizar os passos acima novamente levando em consideração outra combinação de documentos dentro do tempo máximo de 1 hora;
- Responder um questionário de conclusão sobre o processo realizado, mostrado no Apêndice I;

O resultado deste primeiro processo é o *Gold standard* inicial.

O segundo processo desta etapa é idêntico ao processo da primeira etapa, contudo, neste processo, são necessários dois especialistas de domínio, distintos do especialista que realizou o primeiro processo.

O resultado deste segundo processo é *Gold standard* manual.

Após, finalizados os *Gold standard* inicial e *Gold standard* manual é necessário analisar a dispersão dos especialistas, que foi feito diante dos resultados

do *Gold standard* inicial e *Gold standard* manual. Após a análise de dispersão dos especialistas o resultado do *Gold standard* manual é comparado com a análise automática, feita através da abordagem desenvolvida neste trabalho de mestrado. Para essa análise, são utilizadas as métricas de Precisão, Revocação e F-*Measure.*

Segunda Etapa

A segunda etapa deste experimento será feita de forma manual juntamente com os termos da ontologia. A ontologia serve para guiar os especialistas de domínio a fim de encontrar palavras idênticas entre documentos. Para realizar a segunda etapa serão necessários três especialistas e um corpus composto de aproximadamente 24 documentos com documentos diferentes aos analisados na primeira etapa.

Essa etapa é composta por dois processos: O primeiro processo é a criação do *Gold standard* inicial, que é feito por um especialista de domínio. Neste processo, os documentos são analisados em pares, ou seja, o primeiro documento é comparado com todos os demais do *corpus* (exemplo: o primeiro documento com o segundo documento, o primeiro documento com o terceiro, e assim até terminar a lista). Para realizar este processo o especialista de domínio seguirá os seguintes passos:

- Ler o par de documentos escolhido com atenção, mas sem fazer nenhuma anotação;
- Ler os termos presentes na ontologia;
- Ler novamente o par de documentos escolhido anotando as palavras que são termos da ontologia e aparecem em ambos os documentos;
- Ler pela terceira vez o par de documentos a fim de encontrar outras palavras ontológicas idênticas em ambos os documentos;
- Anotar os números dos documentos e a quantidade de palavras idênticas encontradas;
- Realizar os passos acima novamente levando em consideração outra combinação de documentos dentro do tempo máximo de 1 hora;

- Responder um questionário de conclusão sobre o processo realizado, mostrado no Apêndice II;

O resultado deste segundo processo é o *Gold standard* inicial.

O segundo processo desta etapa é idêntico ao processo da primeira etapa, contudo, neste processo, serão necessários dois especialistas de domínio, distintos do especialista que realizou o primeiro processo e também distintos dos especialistas que realizaram a primeira etapa.

O resultado deste segundo processo é *Gold standard* manual com a ontologia como instrumento.

Após, finalizados os *Gold standard* inicial e *Gold standard* manual é necessário analisar a dispersão dos especialistas, que foi feito diante dos resultados do *Gold standard* inicial e *Gold standard* manual.

Após a análise de dispersão dos especialistas o resultado do *Gold standard* manual é comparado com a análise automática, feita através da abordagem desenvolvida neste trabalho de mestrado. Para essa análise, são utilizadas as métricas de Precisão, Revocação e *F-Measure.*

Terceira Etapa

A terceira etapa é chamada de semiautomática, pois é realizada de forma automática e manual. Esta etapa é realizada com todos os especialistas envolvidos nas duas primeiras etapas. O corpus utilizado é composto por documentos diferentes daqueles analisados nas duas primeiras etapas, mas têm estruturas parecidas.

O primeiro processo desta etapa é feito de forma automática, ou seja, o corpus é analisado pela abordagem desenvolvida. A saída deste processo são os resultados de quais documentos do corpus são similares.

Após este processo o corpus processado automaticamente é entregue aos especialistas de domínio juntamente com os resultados obtidos. Os especialistas de domínio fazem a análise manual do corpus, seguindo os resultados obtidos

automaticamente. É dado aos especialistas pares de documentos com diferentes tipos de similaridades: pares com similaridade acima de 85%, pares com similaridade entre 45% e 84%, pares com similaridade entre 5 e 44% e pares com similaridade de 0 a 5%. Além disso, todos os documentos estarão com as palavras ontológicas na cor vermelha. Com estes resultados, os especialistas irão analisar os pares manualmente e obter suas conclusões quanto à similaridade dos pares analisados. Para realizar o processo de análise, os especialistas seguem os seguintes passos:

- Ler o par de documentos escolhido com atenção, mas sem fazer nenhuma marcação;
- Ler novamente o par de documentos escolhido, anotando as palavras ontológicas idênticas nos dois documentos. Para a anotação de palavras idênticas, não são considerados os artigos, os numerais, as conjunções e as preposições.
- Ler pela terceira vez o par de documentos a fim de encontrar palavras idênticas em ambos;
- Anotar os números dos documentos e a quantidade de palavras idênticas encontradas.
- Realizar os passos acima novamente levando em consideração outra combinação de documentos dentro do tempo máximo de 1 hora;
- Responder um questionário de conclusão sobre o processo realizado, mostrado no Apêndice III;

Após a análise e resultados obtidos pelos especialistas, guiados pelos resultados obtidos automaticamente, são utilizadas as métricas de Precisão, Relocação e *F-Measure*, para comparar os resultados obtidos automaticamente com os resultados obtidos pelos especialistas, realizados de forma semiautomáticos. Para as três etapas, os especialistas de domínios são os mesmos.

Análise dos Resultados

Neste tópico são analisados os resultados obtidos nas três etapas do experimento.

Resultados da Primeira Etapa

O Gold Standard inicial da primeira etapa foi feito pelo especialista de domínio que encontrou 23 pares de palavras idênticas entre um par de documentos em um tempo de 60 minutos.

O anotador 1 encontrou 18 pares de palavras idênticas. O anotador 3 encontrou 16 pares de palavras idênticas, todos eles realizando a análise nos mesmos pares de documentos em um tempo de 60 minutos.

Dessa forma é possível notar a dispersão encontrada entre o especialista de domínio e os anotadores. Dessa forma, o *Gold Standard* manual é composto por 23 pares de palavras idênticas encontradas pelos especialistas em um par de documentos dentro de 60 minutos.

Diante do *Gold Standard* manual, as métricas para comparar as anotações feitas nesta primeira etapa com a etapa automática, são as métricas de precisão, revocação e F- Mensure, citadas no capítulo 2.

O total de pares de palavras idênticas encontradas nos documentos pelo algoritmo são 30 pares. O total de pares de palavras não anotadas manualmente e que foram encontradas pelo algoritmo são 7 pares. A Tabela 1 mostra o resultado das medidas de desempenho.

Tabela 1: Medidas de desempenho da primeira etapa

Precisão	Revocação	F- Measure
0,91	0,7	0,74

Fonte: Autor

Esses resultados mostrados na Tabela 1,evidenciam as dificuldades encontradas pelos anotadores na anotação manual dos documentos.

Resultados da Segunda Etapa

O Gold Standard inicial da primeira etapa foi feito pelo especialista de domínio que encontrou 17 pares de palavras idênticas entre um par de documentos em um tempo de 60 minutos.

O anotador 1 encontrou também 17 pares de palavras idênticas. O anotador 3 encontrou 16 pares de palavras idênticas, todos eles realizando a análise nos mesmos pares de documentos em um tempo de 60 minutos.

Dessa forma é possível notar a dispersão encontrada entre o especialista de domínio e os anotadores. Dessa forma, o *Gold Standard* manual é composto por 17 pares de palavras idênticas encontradas pelos especialistas em um par de documentos dentro de 60 minutos.

Diante do *Gold Standard* manual, as métricas para comparar as anotações feitas nesta primeira etapa com a etapa automática, são as métricas de precisão, revocação e F- Mensure, citadas no capítulo 2.

O total de pares de palavras idênticas encontrados nos documentos pelo algoritmo são 21 pares. O total de pares de palavras não anotadas manualmente e que foram encontradas pelo algoritmo são 4 pares. A Tabela 2 mostra o resultado das medidas de desempenho para a segunda etapa.

Tabela 2: Medidas de desempenho da segunda etapa

Precisão	Revocação	F- Measure
0,94	0,76	0,83

Fonte: Autor

Os resultados mostrados na Tabela 2, que com o auxílio da ontologia os anotadores obtiveram dificuldades para encontrar palavras idênticas na ontologia, contudo, comparando a Tabela 2 com a Tabela 1, no primeiro experimento, é possível notar que os resultados das medidas aumentaram consideravelmente.

Resultados da Terceira Etapa

O Gold Standard inicial da primeira etapa foi feito pelo especialista de domínio que encontrou 28 pares palavras idênticas entre um par de documentos em um tempo de 60 minutos.

O anotador 1 encontrou também 28 pares de palavras idênticas. O anotador 3 encontrou 29 pares de palavras idênticas, todos eles realizando a análise nos mesmos pares de documentos em um tempo de 60 minutos.

Dessa forma é possível notar a dispersão encontrada entre o especialista de domínio e os anotadores. Dessa forma, o *Gold Standard* manual é composto por 28 pares de palavras idênticas encontradas pelos especialistas em um par de documentos dentro de 60 minutos.

Diante do *Gold Standard* manual, as métricas para comparar as anotações feitas nesta primeira etapa com a etapa automática, são as métricas de precisão, revocação e F- Mensure, citadas no capítulo 2.

O total de pares de palavras idênticas encontrados nos documentos pelo algoritmo são 29 pares. O total de pares de palavras não anotadas manualmente e que foram encontradas pelo algoritmo é um par. A Tabela 3 mostra o resultado das medidas de desempenho para a segunda etapa.

Tabela 3: Medidas de desempenho da terceira etapa

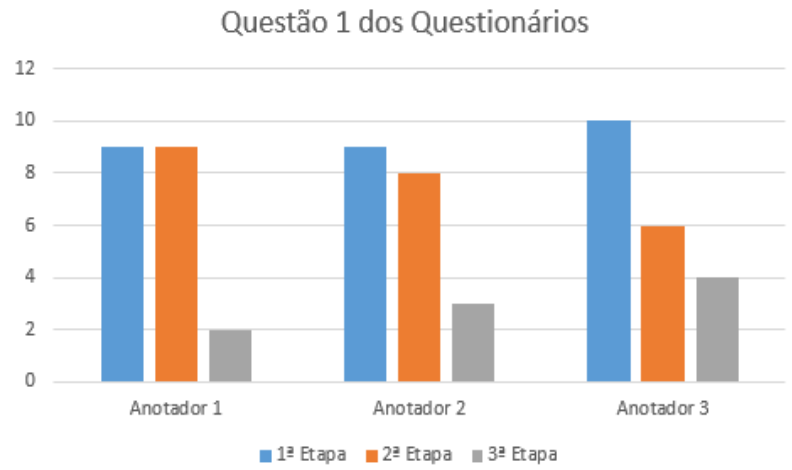
Precisão	Revocação	F- Measure
0,96	0,93	0,94

Fonte: Autor

Respostas dos questionários nas três etapas (Apêndice A, B e C)

Questão 1: Em uma escala de 0 a 10, onde 0 significa pouca dificuldade e 10 significa extrema dificuldade, indique o grau de dificuldade que você teve para encontrar palavras idênticas em pares de documentos.

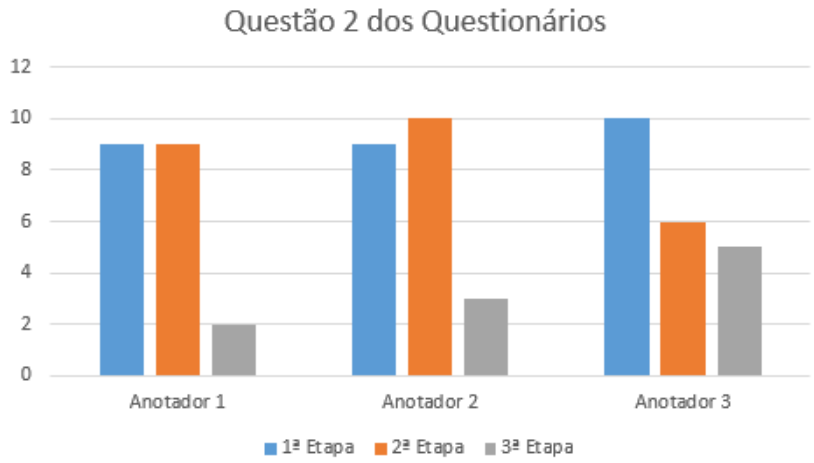
Gráfico 1: Grau de Dificuldade percebido para encontrar palavras idênticas nos documentos.



Fonte: Autor

Questão 2: Em uma escala de 0 a 10, onde 0 significa pouco cansaço e 10 significa extremo cansaço, indique o grau de cansaço que você sentiu para encontrar palavras idênticas em pares de documentos.

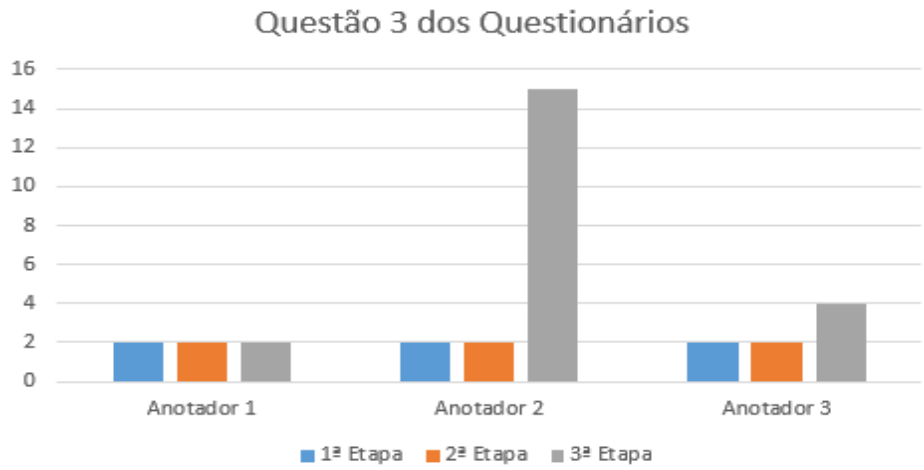
Gráfico 2: Grau de Cansaço percebido para encontrar palavras idênticas.



Fonte: Autor

Questão 3: Em uma escala de 0 a 20, onde 0 significa análise de nenhum documento e 20 significa análise de mais de 20 documentos, indique a quantidade de documentos que você conseguiu analisar no tempo máximo de 60 minutos.

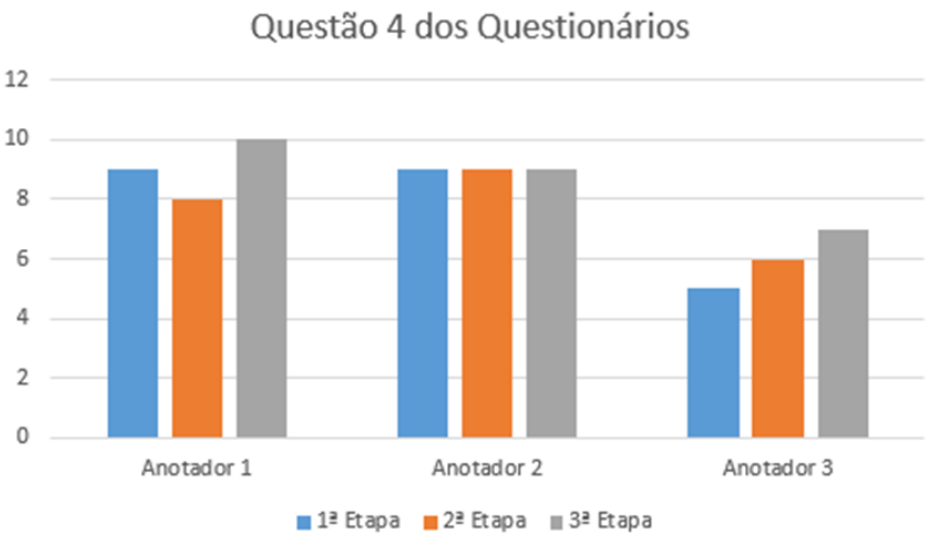
Gráfico 3: Quantidade de documentos analisados pelos anotadores.



Fonte: Autor

Questão 4: Em uma escala de 0 a 10, onde 0 significa pouca eficácia e 10 significa eficácia extrema, indique o grau de eficácia na obtenção de palavras idênticas entre os documentos.

Gráfico 4: Grau de Eficácia percebido para encontrar palavras idênticas.



Fonte: Autor

Na primeira etapa, as análises dos documentos foram feitas de forma manual, conforme descrição do experimento. Conforme Gráfico 1, os anotadores têm extrema dificuldade para encontrar palavras idênticas nos documentos, tornando a análise dos mesmos muito cansativa, como mostrado no Gráfico 2. Todos os anotadores conseguiram analisar apenas 2 documentos, ou seja, um par de documentos, em um tempo de 60 minutos, conforme Gráfico 3. Este fato é resultado da dificuldade em encontrar palavras idênticas nos documentos e do cansaço que esta tarefa gera aos anotadores. Apesar de toda a dificuldade e cansaço para analisar os documentos, dois anotadores consideram que a análise dos documentos foi feita de forma muito eficaz. Apenas um anotador considerou sua anotação não muito eficaz, conforme Gráfico 4.

As dificuldades relatadas pelos anotadores já eram esperadas devido ao texto no português informal encontrado nas entrevistas.

Na segunda e na terceira etapa, os anotadores realizaram os mesmos passos que a primeira etapa, contudo, os documentos são distintos nas três etapas. Na etapa 2, os anotadores têm como apoio os termos ontológicos e na etapa 3 os anotadores tiveram o auxílio da abordagem SimIGroup.

O Gráfico 1 mostra o grau de dificuldade obtido por cada anotador para detectar as palavras idênticas nos documentos. Observando as 3 etapas é possível identificar, no geral, que as dificuldades dos anotadores diminuíram com o auxílio da ontologia, utilizada na segunda etapa. Observando a terceira etapa, nota-se que o auxílio da abordagem SimIGroup diminui drasticamente a dificuldade dos anotadores na análise dos documentos para obter palavras idênticas em relação as etapas anteriores.

No Gráfico 2, observando a primeira e segunda etapa, percebe-se que houve uma queda do grau de cansaço dos anotadores com o auxílio da ontologia na realização desta tarefa. Na terceira etapa, com o auxílio da abordagem SimIGroup, percebe-se que o grau de cansaço diminuiu muito com o auxílio da abordagem SimIGroup.

No Gráfico 3, que mostra a quantidade de documentos analisados, nota-se que não houve mudanças na quantidade de documentos estudados com o auxílio da ontologia. Com o auxílio da abordagem SimIGroup, na terceira etapa, é possível perceber que houve um aumento de documentos analisados.

No Gráfico 4, que mostra o grau de eficácia na obtenção de palavras idênticas com o auxílio da ontologia para os anotadores, observando as duas primeiras etapas, percebe-se que não houve grande alteração dos pesquisadores no grau de eficácia com o auxílio da ontologia. Também é possível perceber que o grau de eficácia na obtenção de palavras idênticas aumentou consideravelmente com o auxílio da abordagem SimIGroup.

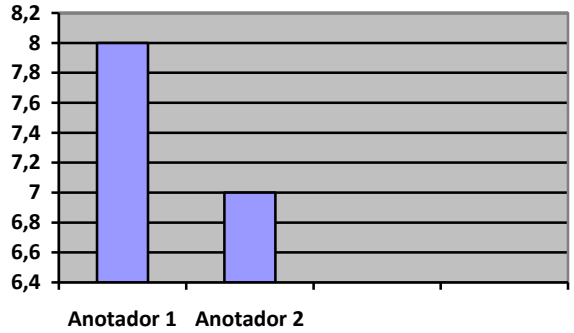
Diante dos resultados desta terceira etapa, nota-se que o uso da abordagem SimIGroup, auxilia, de maneira positiva, os anotadores na análise de documentos, diminuindo as dificuldades, o cansaço e aumentando o número de documentos analisados e a eficácia da análise.

Respostas do questionário da Terceira Etapa (Apêndice D)

O questionário final analisa, por parte dos anotadores, se o método desenvolvido auxilia na análise qualitativa de dados, usando a técnica *Coding*. O Gráficos 5, 6 e 7 mostram as respostas dos anotadores/especialistas de domínio. O anotador 3, por desconhecer a técnica de *Coding*, não realizou o questionário final.

Questão 1: Em uma escala de 0 a 10, onde 0 significa pouca facilidade e 10 significa extrema facilidade, indique o grau de facilidade que você tem para encontrar *codes* após a análise dos documentos feita na terceira etapa.

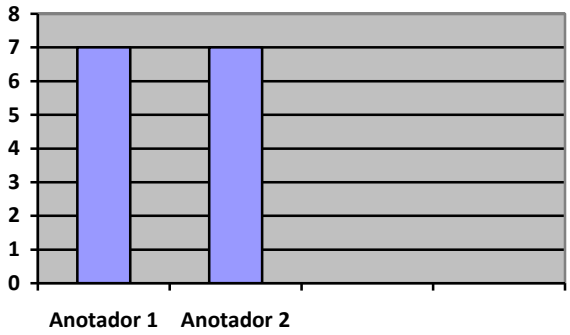
Gráfico 5: Grau de facilidade para encontrar *codes* nos documentos com a abordagem SimIGroup



Fonte: Autor

Questão 2: Em uma escala de 0 a 10, onde 0 significa pouca facilidade e 10 significa extrema facilidade, indique o grau de facilidade que você tem para encontrar *quotations* após a análise dos documentos feita na terceira etapa.

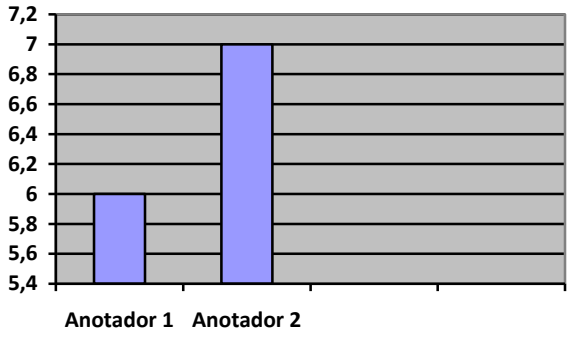
Gráfico 6: Grau de facilidade para encontrar *quotations* nos documentos com a abordagem SimIGroup



Fonte: Autor

Questão 3: Indique, numa escala de 0 a 10, o grau de facilidade em analisar qualitativamente os documentos, onde 0 significa pouca facilidade e 10 significa facilidade extrema, a partir do trabalho produzido.

Gráfico 7: Grau de facilidade para analisar qualitativamente os documentos com o auxílio da abordagem SimIGroup



Fonte: Autor

Nos Gráficos 5 e 6, os pesquisadores mostraram um grau positivo sobre a facilidade de encontrar *codes* e *quotations* nos documentos com o auxílio da abordagem SimIGroup. Para encontrar o *codes*, o grau de facilidade indicado foi

maior do que para encontrar os *quotations*. Este fato já era esperado, pois os *quotations* são mais complexos do que os *codes*. No Gráfico 7, os pesquisadores indicaram o grau de facilidade para analisar qualitativamente os documentos. Este grau também é positivo, levando em consideração a complexidade da análise qualitativa em documentos, principalmente os documentos em estudo, por conter texto com linguagem informal.

Diante dos experimentos realizados e as repostas aos questionários, fica evidente que o método desenvolvido, SOM4SIMD, juntamente com a abordagem SimIGroup auxilia os pesquisadores na análise de documentos, especialmente na análise qualitativa de dados.

5.2. Comparação do Método SOM4SIMD com os trabalhos correlatos

Nesta sessão é feita a comparação do método desenvolvido com os trabalhos correlatos citados no capítulo 2. Essa comparação permite mostrar as vantagens do método SOM4SIMD em comparação com os métodos mostrados.

A Tabela 4 mostra um comparativo entre o trabalho desenvolvido com os trabalhos correlatos.

Tabela 4: Comparação do método desenvolvido com os trabalhos correlatos

Autores	Tipos de Texto	Artefato utilizado na obtenção de similaridade	Modo de utilização do Artefato	Resultados
Liu e Wang (2014)	Textos formais em inglês	Ontologia	Hierarquia da ontologia	Precisão: 0,71 Revocação: 0,91 Medida F: 0,80
Jingling, Huiyun e Baojiang (2014)	Textos formais em inglês	Vetor de palavras	Semântica das palavras e ordem das mesmas na sentença	Precisão: - Revocação: - Medida F: - Acurácia: 98,5

Gupta e Yadav (2014)	Textos formais em inglês	Base dados Word Net	Hierarquia da base de dados	Precisão: - Revocação: - Medida F: - Acurácia: -
SOM4SIMD	Textos informais em português Textos semiestruturado Texto com dialetos regionais	Ontologia	Hierarquia da ontologia Variantes e sinônimos e Relacionamentos	Precisão: 0,96 Revocação: 0,93 Medida F: 0,94 Acurácia: 96,0

Fonte: Autor

O primeiro trabalho correlato foi o trabalho de Liu e Wang (2014), que desenvolveram um método de similaridade semântica entre sentenças e textos grandes.

De acordo com a Tabela 4, o método SOM4SIMD, contém todas as medidas de desempenho melhor do que o método desenvolvido por Liu e Wang (2014). Este fato mostra que, utilizar os termos formais presentes na ontologia, juntamente com seus sinônimos e variantes, aumenta a eficácia na obtenção de similaridade semântica entre documentos. O trabalho de Liu e Wang (2014), não usa os sinônimos e variantes dos termos formais da ontologia, apenas a estrutura hierárquica vertical e horizontal. Estes autores também não fazem a comparação de relacionamentos entre documentos, o que aumenta a similaridade dos mesmos.

Jingling, Huiyun e Baojiang (2014) desenvolveram um trabalho que encontra a similaridade semântica entre sentenças, utilizando vetor de palavras. O método proposto pelos autores é aplicável a textos pequenos, levando em consideração informações semânticas, estruturais e de ordem das palavras em uma sentença. Os autores utilizaram a medida de desempenho Acurácia. A fórmula da acurácia utilizada foi dividir o número de resultados corretos pelo número de sentenças testadas multiplicado por 100%. A acurácia para o método proposto é 98,5. Para fazer a acurácia no método SOM4SIMD, idêntica com a fórmula proposta por Jingling, Huiyun e Baojiang (2014), foi levada em consideração a anotação dos especialistas, onde o número de anotações corretas foi 28, dividido por 29, que o número de total de anotações. O resultado da 96, 5, abaixo do resultado obtido

pelos autores que foi 98,5. Este trabalho não utiliza ontologia para encontrar as palavras idênticas, diferindo bastante do método desenvolvido neste trabalho de mestrado. Com o uso da ontologia de domínio, a similaridade torna-se mais fácil e mais eficaz na obtenção de termos chaves do domínio

No trabalho de Gupta e Yadav (2014) é proposta a criação de uma métrica para o cálculo da relação semântica entre pares de conceitos, utilizando a abordagem baseada em características de Tversky que leva em consideração a característica comum e distinta dos dois termos ou conceitos. Se a semelhança é maior comparada com as diferenças, a similaridade entre os conceitos é alta, caso contrário a similaridade é baixa. A avaliação dessa abordagem leva em consideração o método de classificação chamado de Rubenstein Goodenough, que é um método onde é considerado o julgamento humano, ou seja, o quanto um especialista julga o quanto um par de conceito é idêntico. Esses pares de conceitos foram extraídos de uma base de dados chamada Word Net. Dessa forma, comparar o método dos autores com o método SOM4SIMD é uma tarefa difícil, pois a base de dados aqui utilizada é uma ontologia do próprio domínio, não tem a estrutura parecida com a Word Net. No trabalho de Gupta e Yadav (2014), também não é levado em consideração, os similares e variantes da ontologia, e nem tão pouco os relacionamentos ontológicos presentes nos documentos no momento da comparação.

5.3. Contribuições e Limitações

O método SOM4SImD resultou na similaridade semântica entre documentos mais eficientes do que os demais métodos presentes na literatura, sendo uma das contribuições científicas para a área de Ciência da Computação. Isto porque o SOM4SImD considera os sinônimos e variantes dos termos presentes na ontologia. As demais medidas de similaridades desenvolvidas consideram somente a hierarquia horizontal e vertical da ontologia. Este fato torna mais eficiente no cálculo da similaridade, pois palavras sinônimas são consideradas similares.

Outra contribuição deste trabalho foi o desenvolvimento de uma abordagem que auxilia os pesquisadores da área da educação especial, na análise qualitativa

de dados dos documentos em formato de entrevistas. A abordagem mostra os documentos similares em uma rede de visualização semântica, com cores que indicam porcentagem de similaridade entre eles. Ao clicar em qualquer documento, este é aberto com os termos ontológicos em vermelho, mostrando aos pesquisadores que estes termos são importantes naquele documento e, portanto, são partes de marcações realizadas na técnica *Coding* da análise qualitativa.

Existem limitações neste trabalho de mestrado. A primeira delas é o fato da ontologia ter sido representada em uma estrutura interna na implementação do método. Caso esta ontologia cresça muito, torna-se complicado a representação desta em memória. Representar a ontologia internamente, também tornou o processamento dos algoritmos um pouco lento, tornando mais demorado este processamento.

Outra questão é uma limitação deste trabalho desenvolvido e que são consideradas apenas as palavras e relações ontológicas dos documentos para comparação de similaridade, deixando de lado o significado semântico do contexto de uma frase ou parágrafo.

Contudo, os resultados dos experimentos mostram que mesmo com as limitações, o trabalho desenvolvido trouxe benefícios no momento de obtenção da similaridade e no auxílio das análises para os anotadores/ especialistas.

Capítulo 6

CONCLUSÕES E TRABALHOS FUTUROS

Nesta dissertação de mestrado foi desenvolvido um método que detecta similaridade semântica entre documentos compostos por entrevistas sobre a educação especial, com uma linguagem informal do português brasileiro. Juntamente com este método, criou-se uma abordagem que auxilia pesquisadores na análise de documentos, principalmente a análise qualitativa.

Diante das pesquisas bibliográficas realizadas, observou-se que nenhuma das abordagens criadas contemplam o português informal brasileiro, apenas línguas mais formais, como o inglês e o mandarim. Para criar o método desenvolvido nesta dissertação, foram utilizados vários artefatos já existentes. Um destes artefatos foi a ontologia de domínio Educação Especial, que contempla o mesmo domínio dos documentos utilizados como estudo de caso. A utilização desta ontologia foi de extrema importância, pois foi através do enriquecimento dela, com termos sinônimos e variantes, que foi possível a obtenção de maior grau de similaridade entre documentos do que os demais trabalhos encontrados na literatura.

Apesar das limitações apresentadas no capítulo anterior, os resultados dos experimentos mostram que a utilização do método SOM4SImD traz maior grau de similaridade entre os documentos e, juntamente com a abordagem SimIGroup auxilia

pesquisadores na análise de documentos, essencialmente na análise qualitativa de dados, quando utilizada a técnica de *Coding*.

Futuramente, outros trabalhos podem ser desenvolvidos a partir dessa dissertação. Um desses trabalhos é desenvolver um algoritmo que aumente o desempenho de processamento dos documentos em relação à ontologia, pois ela encontra-se em uma estrutura interna, o que é limitação discutida no capítulo anterior. A ontologia poderá ser retirada da estrutura interna, sendo consultada somente nos momentos necessários, fazendo assim aumentar ainda mais o desempenho do método.

Outro trabalho futuro que poderá ser desenvolvido é o aperfeiçoamento da abordagem SimIGroup. Poderá ser desenvolvido a partir dessa abordagem uma ferramenta, em que os especialistas de domínio consigam realizar a análise qualitativa de forma interativa e completa dos documentos, utilizando a técnica *Coding*.

REFERÊNCIAS

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001.

BORST, W. N. *Construction of engineering ontologies for knowledge sharing and reuse*. Universiteit Twente, 1997.

COWIE, J.; LEHNERT, W. Information extraction. *Communications of the ACM*, ACM, v. 39, n. 1, p. 80–91, 1996.

FERNANDES, W. L.. *Desenvolvimento de uma Solução Computacional para Análise de Dados Qualitativos na Peerspectiva da Educação Especial*. PHD thesis, PPGEEs, UFSCAR, São Carlos, SP, 2015.

FRAKES, W. B.; BAEZA-YATES, R. *Information retrieval: data structures and algorithms*. Prentice Hall PTR, 1992.

GRUBER, T. A translation approach to portable ontology specifications. *Knowledge acquisition, Elsevier*, v. 5, n. 2, p. 199–220, 1993.

GRUBER, T. Ontology. *Encyclopedia of database systems*, Springer, p. 1963–1965, 2009.

GUPTA e D. K. YADAV. Semantic Similarity Measure Using Information Content Approach With Depth For Similarity Calculation*. In International Journal of Scientific and Technology Research*, vol. 3, p. 165-169, 2014.

GUARINO, N. Formal ontology in information systems: *proceedings of the first international conference (FOIS'98)*, June 6-8, Trento, Italy. [S.l.]: Ios PressInc, 1998.

HANCOCK, B. *Trent Focus for Research and Development in Primary Health Care: An Introduction to Qualitative Research*. Trent Focus, 2002.

HERNANDES, E. C. M. *Apoio à condução de análise qualitativa com técnicas de visualização e mineração de texto*. Qualificação (doutorado) – Programação de Pós Graduação em Ciência da Computação. Departamento de Computação, Universidade Federal de São Carlos, 2012.

KIRYAKOV, A. et al. Semantic annotation, indexing, and retrieval. *Web Semantics: Science, Services and Agents on the World Wide Web*, Elsevier, v. 2, n. 1, p. 49–79, 2004.

LI, Y.; BONTCHEVA, K. Hierarchical, perceptron-like learning for ontology-based information extraction. *In: ACM. Proceedings of the 16th international conference on World Wide Web*. [S.l.], 2007. p. 777–786.

LIN, D. An information-theoretic definition of similarity. *In Proceedings of the 15th International Conference on Machine Learning,Madison*, WI, 1998.

WU, PALMER M.. Verbs semantics and lexical selection. *In Paper presented at the proceedings of the 32nd annual meeting on association for computational linguistics*, 1994.

LIU, H.; WANG, P.. Assessing Text Semantic Similarity Using Ontology. *In JOURNAL OF SOFTWARE*, Vol. 9, N. 2, p. 490-497, February 2014.

MANNING, C. D., SCHÜTZE H. (1999). *Foundations of statistical natural language processing* (Ed.). MIT press.

MAYNARD, D.; PETERS, W.; LI, Y. Metrics for evaluation of ontology-based information extraction. In: EDINBURGH, UK. *International world wide web conference*. [S.l.], 2006.

NOY, N. F.; MCGUINNESS, D. L. *Ontology Development 101: A Guide to Creating Your First Ontology*. [S.l.], 2001.

PEREIRA, J. W. *Anotação semântica baseada em ontologia: um estudo do português brasileiro em documentos históricos do final do século XIX*. Dissertação (mestrado). Universidade Federal de São Carlos, 2014.

RESNIK, P. Using information content to evaluate semantic similarity in a taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Montréall; 448-53, 1995.

SEAMAN, C.B. Qualitative methods in empirical studies of software engineering. *IEEE Transactions on Software Engineering*, Los Alamitos, v. 25, n. 4, p.557-573, Jul./Aug. 1999.

SUSSNA, M. Word sense disambiguation for free-test indexing using a massive semantic network. *Proceedings of the 2nd International Conference on Information and Knowledge Management*. Arlington, Virginia, USA, 1993.

STRAUSS A, CORBIN, J., *Basics of qualitative research: techniques and procedures for developing grounded theory***.** Sage Publications, 3ed, 2008.

STUDER, R.; BENJAMINS, V. R.; FENSEL, D. Knowledge engineering: principles and methods. *Data & knowledge engineering, Elsevier*, v. 25, n. 1, p. 161–197, 1998.

UREN, V. et al. Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: science, services and agents on the World Wide Web*, Elsevier, v. 4, n. 1, p. 14–28, 2006.

YAGUINUMA, C. A.; SANTOS, M. T.; BIAJIZ, M. Meta-ontologia difusa para representação de informações imprecisas em ontologias. In: *Workshop on Ontologies and Metamodeling in Software and Data Engineering*. [S.l.: s.n.], 2007. p. 57–67.

WIMALASURIYA, D. C.; DOU, D. Components for information extraction: ontology-based information extractors and generic platforms. *In: ACM. Proceedings of the 19th ACM international conference on Information and knowledge management*. [S.l.], 2010. p. 9–18.

ZHAO, J.; ZHANG, H.; CUI B. Sentence Similarity Based on Semantic Vector Model. In *3PGCIC '14 Proceedings of the 2014 Ninth International Conference on P2P, Parallel, Grid, Cloud and Internet Computing,* P. 499-503, 2014.

Apêndice A*

QUESTIONÁRIO DA PRIMEIRA ETAPA DOS EXPERIMENTOS

Questionário para a primeira e segunda etapa dos experimentos

1) Em uma escala de 0 a 10, onde 0 significa pouca dificuldade e 10 significa extrema dificuldade, indique o grau de dificuldade que você teve para encontrar palavras idênticas em pares de documentos.

2) Em uma escala de 0 a 10, onde 0 significa pouco cansaço e 10 significa extremo cansaço, indique o grau de cansaço que você sentiu ao realizar a tarefa proposta.

3) Em uma escala de 0 a 20, onde 0 significa análise de nenhum documento e 20 significa análise de mais de 20 documentos, indique a quantidade de documentos que você conseguiu analisar no tempo máximo de 60 minutos.

4) Em uma escala de 0 a 10, onde 0 significa pouca eficácia e 10 significa eficácia extrema, indique o grau de eficácia na obtenção de palavras idênticas entre os documentos;

Apêndice B*

QUESTIONÁRIO DA SEGUNDA ETAPA DOS EXPERIMENTOS

Questionário para a segunda e segunda etapa dos experimentos

1) Em uma escala de 0 a 10, onde 0 significa pouca dificuldade e 10 significa extrema dificuldade, indique o grau de dificuldade que você teve para encontrar palavras idênticas em pares de documentos.

2) Em uma escala de 0 a 10, onde 0 significa pouco cansaço e 10 significa extremo cansaço, indique o grau de cansaço que você sentiu ao realizar a tarefa proposta.

3) Em uma escala de 0 a 20, onde 0 significa análise de nenhum documento e 20 significa análise de mais de 20 documentos, indique a quantidade de documentos que você conseguiu analisar no tempo máximo de 60 minutos.

4) Em uma escala de 0 a 10, onde 0 significa pouca eficácia e 10 significa eficácia extrema, indique o grau de eficácia na obtenção de palavras idênticas entre os documentos;

Apêndice C*

QUESTIONÁRIO DA TERCEIRA ETAPA DOS EXPERIMENTOS

Questionário para a segunda e segunda etapa dos experimentos

1) Em uma escala de 0 a 10, onde 0 significa pouca dificuldade e 10 significa extrema dificuldade, indique o grau de dificuldade que você teve para encontrar palavras idênticas em pares de documentos.

2) Em uma escala de 0 a 10, onde 0 significa pouco cansaço e 10 significa extremo cansaço, indique o grau de cansaço que você sentiu ao realizar a tarefa proposta.

3) Em uma escala de 0 a 20, onde 0 significa análise de nenhum documento e 20 significa análise de mais de 20 documentos, indique a quantidade de documentos que você conseguiu analisar no tempo máximo de 60 minutos.

4) Em uma escala de 0 a 10, onde 0 significa pouca eficácia e 10 significa eficácia extrema, indique o grau de eficácia na obtenção de palavras idênticas entre os documentos;

Apêndice D*

QUESTIONÁRIO FINAL DOS EXPERIMENTOS

1) Em uma escala de 0 a 10, onde 0 significa pouca facilidade e 10 significa extrema facilidade, indique o grau de facilidade que você tem para encontrar *codes* após a análise dos documentos feita na terceira etapa.

2) Em uma escala de 0 a 10, onde 0 significa pouca facilidade e 10 significa extrema facilidade, indique o grau de facilidade que você tem para encontrar *quotations* após a análise dos documentos feita na terceira etapa.

3) Indique, numa escala de 0 a 10, o grau de facilidade em analisar qualitativamente os documentos, onde 0 significa pouca facilidade e 10 significa facilidade extrema, a partir do trabalho produzido.