UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

UM MÉTODO ADAPTADO PARA IDENTIFICAÇÃO AUTOMÁTICA DE PRÉ-REQUISITOS ENTRE CONCEITOS EM UM GRAFO DE CONHECIMENTO

Rúbia Silveira de Almeida

Orientador
Sean Wolfgand Matsui Siqueira

RIO DE JANEIRO, RJ - BRASIL
ABRIL DE 2020

RÚBIA SILVEIRA DE ALMEIDA

UM MÉTODO ADAPTADO PARA IDENTIFICAÇÃO AUTOMÁTICA DE PRÉ-REQUISITOS ENTRE CONCEITOS EM UM GRAFO DE CONHECIMENTO

Natureza do trabalho apresentada ao Programa de Pós-Graduação em Informática da Universidade Federal do Estado do Rio de Janeiro, como pré-requisito para a obtenção do grau de Mestre em Informática.

Orientação:
Sean Wolfgand Matsui Siqueira

Rio de Janeiro

2020

UM MÉTODO ADAPTADO PARA IDENTIFICAÇÃO AUTOMÁTICA DE PRÉ-REQUISITOS ENTRE CONCEITOS EM UM GRAFO DE CONHECIMENTO

Rúbia Silveira de Almeida

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓSGRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Sean Wolfgand Matsui Siqueira, D.Sc – UNIRIO

Edmar Welington Oliveira, D.Sc – UFJF

Tadeu Moreira de Classe, D.Sc – UNIRIO

RIO DE JANEIRO, RJ - BRASIL

ABRIL DE 2020

Catalogação informatizada pelo(a) autor(a)

A447 Almeida, Rúbia Silveira de
Um Método Adaptado para Identificação Automática de Pré-Requisitos entre Conceitos em um Grafo de Conhecimento / Rúbia Silveira de Almeida. -- Rio de Janeiro, 2020.
88

Orientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2020.

1. Pré-Requisito entre Conceitos. 2. DBpedia. 3. Dados Conectados. 4. Grafo de Conhecimento. I. Siqueira, Sean Wolfgand Matsui, orient. II. Título.

Este trabalho é dedicado às pessoas que acreditam que "nunca é tarde demais para ser aquilo que você deveria ser (by George Eliot)", pois tarde foi ontem, quando deixamos de tentar.

AGRADECIMENTOS

Primeiramente a Deus e ao Universo que trabalharam a meu favor me proporcionando inspiração para concluir este trabalho.

Agradeço imensamente aos meus pais por me conduzirem ao caminho do conhecimento. Sempre me incentivaram a ir além, especialmente minha mãe que por diversas ocasiões acreditou em minha capacidade e esteve presente me acolhendo em inúmeras noites em claro de dedicação ao longo da vida.

À minha avó Olinda (in memoriam), cuja presença foi essencial na minha vida. Ela sempre nos presenteou com a sabedoria típica dos mais velhos e nos ensinou a sermos resilientes quando necessário.

Em especial agradeço ao professor Sean, meu orientador, com quem compartilhei minhas dúvidas e angústias a respeito do tema. Pelos ensinamentos e todo acompanhamento, além da generosidade, compreensão e paciência no decorrer da minha pesquisa.

Aos colegas pesquisadores professor Bernardo Nunes e professor Ruben Manrique, que me proporcionaram várias ideias para desenvolvimento deste tema de pesquisa.

Agradeço à minha família e aos meus amigos, pela compreensão com os meus diversos momentos de ausência e reclusão durante os estudos.

Aos meus afilhados de casamento Aline e Felipe que foram os primeiros a me incentivarem a ingressar num curso de mestrado e sempre estiveram presentes, torcendo para meu sucesso.

Aos colegas do grupo de estudo do mestrado, Ana Camello, Crystiam Kelle, Davi Faisca, Jerry Medeiros, João Paulo, Marcelo Tibau e Natália Oliveira, pelas ótimas ideias, discussões, conversas e apoio durante a jornada.

Aos colegas de turma André Farzat, Eduardo Augusto, Jackson Queiroz, Solange Santolin e Vanessa Martins que contribuíram em diversas ocasiões com trabalhos em grupo e discussões pertinentes, ampliando nosso conhecimento.

Aos professores do PPGI que tão bem me prepararam para a realização desta dissertação.

Aos funcionários da secretaria do PPGI que sempre muito bem me atenderam quando precisei.

Aos membros da Banca Examinadora pelo trabalho de avaliação.

ALMEIDA, Rúbia Silveira de. **Um Método Adaptado para Identificação Automática de Pré-Requisitos entre Conceitos em um Grafo de Conhecimento.** UNIRIO, 2020. 88 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Com a maior disponibilidade de recursos de aprendizagem online, alunos buscam acessar materiais relevantes na Internet para percorrer um caminho de conteúdos que os conduzirão a aprendizagem. O problema de determinar um caminho de aprendizagem efetivo a partir de recursos de aprendizagem da Web depende da identificação das relações de pré-requisitos entre conceitos nesses recursos. Nesta pesquisa, propõe-se o desenvolvimento de um novo método baseado em uma métrica da literatura para identificação automática de relações de pré-requisito entre conceitos. O método proposto, OP-RefD – Optimized Performance RefD, adapta o RefD para aproveitar o conhecimento extraído dos dados conectados abertos, mais especificamente o grafo de conhecimento DBpedia. Um estudo experimental foi conduzido para avaliar e comparar o método adaptado proposto com a métrica de referência. Os pares conceituais avaliados são submetidos a seus correspondentes na DBpedia, de forma que os diversos relacionamentos existentes entre cada conceito possam contribuir para a avaliação. Os resultados mostram que a estrutura da DBpedia pode ser usada para identificar relações de pré-requisito entre conceitos, sem perda de eficácia, e com um melhor desempenho em relação ao tempo de processamento em referência à base de comparação.

Palavras-chave: Pré-Requisito entre Conceitos, DBpedia, Dados Conectados, Grafo de Conhecimento.

ABSTRACT

With the increased availability of online learning resources, learners search for relevant materials on the Internet in order to follow a content path that leads to learning. The problem of determining an effective learning path from Web resources depends on identifying the prerequisite relationships between concepts in those resources. In this research, a new method based on a metric of the literature is proposed for automatic identification of prerequisite relationships between concepts. The proposed method, OP-RefD – Optimized Performance RefD, adapts the RefD to take advantage of the knowledge extracted from open linked data, more specifically from DBpedia knowledge graph. An experimental study was conducted to evaluate and compare the proposed adapted method with the baseline metric. The conceptual pairs evaluated are submitted to their correspondents in DBpedia, so that the different relationships between each concept can contribute to the evaluation. The results show that DBpedia structure can be used to identify prerequisite relationships between concepts, without loss of effectiveness, and with a better performance of processing time compared to the baseline.

Keywords: Concepts Prerequisite**,** DBpedia, Linked Data, Knowledge Graph.

SUMÁRIO

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

Lista de Siglas

AUC	*Area under the curve*
AWS	*Amazon Web Services*
EQUAL	Função de Peso que considera pesos iguais
LO	*Learning Objects*
MOOC	*Massive Open Online Courses*
NEC	*Named Entities Classification*
OP-RefD	*Optimized Performance Reference Distance*
PRC	*Precision-Recall Curve*
PREREQ	Método para prever pré-requisitos de conceito desconhecido
RDF	*Resources Description Framework*
RefD	*Reference Distance*
ROC	*Receiver Operating Characteristic Curve*
SPARQL	*Query Language for RDF*
TFIDF	*Term Frequency–Inverse Document Frequency*
URI	*Uniform Resource Identifier*
XML	*Extensible Markup Language*
W3C	*World Wide Web Consortium*

1. INTRODUÇÃO

Neste capítulo serão apresentados os elementos que motivaram a realização deste trabalho. O capítulo tem como objetivos apresentar o problema a ser abordado, levantar a hipótese a ser investigada e mostrar as questões de pesquisa que serão exploradas no desenvolvimento desta dissertação.

1.1. Contexto de Pesquisa

A disponibilidade de uma variedade de conteúdos de mídia eletrônica deu origem a novos paradigmas de aprendizagem e de conhecimento (ROY et al., 2008), (KIVUNJA, 2015) e (GOLDIE, 2016). O *e-learning* representa todas as formas de aprendizagem por meio da tecnologia, por meio da Internet.

A adoção do *e-learning* também se beneficiou da iniciativa da Web 3.0 – a Web Semântica. Os primeiros conceitos de Semântica na Web foram apresentados por Tim Berners-Lee et al. (2001): "a Web Semântica não é uma Web separada, mas sim uma extensão da atual, na qual informações ganham um significado bem definido, permitindo que computadores e pessoas trabalhem melhor em cooperação".

Uma das vantagens da Web Semântica para este ambiente de aprendizado online, é que ela permite a representação do conteúdo da Web de uma forma que é mais facilmente processável por máquina e o uso de técnicas inteligentes para aproveitamento dessas representações através do uso de metadados (ANTONIOU; HARMELEN, 2008).

É importante ressaltar que, no cenário de alunos que tentam acessar materiais instrucionais personalizados na Internet é importante mencionar que, pela pobre qualidade de anotação de metadados destes materiais, eles são retornados em ferramentas de busca de forma imprecisa. Então são recuperados como um conjunto de documentos com potencial de serem interessantes, porém podem não ser relevantes para o problema de aprendizagem que o aluno busca resolver.

Com esta natureza disjuntiva dos resultados fornecidos pelos mecanismos de pesquisa tradicionais, torna-se crucial fornecer aos alunos percursos de aprendizagem adaptados que proponham uma sequência de recursos que correspondam aos seus objetivos de aprendizagem. Desta forma, a sequência deve ser tal que cada novo documento se baseie em conceitos que já foram definidos em documentos anteriores (CHANGUEL et al., 2015).

Para Ausubel (1963), o aluno consegue desenvolver um aprendizado significativo de novos conteúdos quando há vinculação com os conceitos que o mesmo já conhece, já estão incorporados. O desafio não é apenas tornar os materiais facilmente disponíveis na Web, mas torná-los utilizáveis de uma maneira que satisfaçam os objetivos de aprendizagem específicos de um determinado aluno que segue um determinado curso.

O problema de determinar um caminho de aprendizagem efetivo a partir de conteúdos de aprendizagem da Web depende da identificação precisa do resultado, dos conceitos de pré-requisito nesses documentos e de sua ordenação de acordo com essas informações. No contexto desta dissertação, uma relação de pré-requisito descreve uma relação básica entre conceitos em cognição, educação e em outras áreas (LIANG et al., 2015).

Um pré-requisito geralmente é um conceito ou requerimento no qual se exige conhecimento prévio antes que se possa aprender o próximo conceito. Assim, quando

uma pessoa está apreendendo, organizando, aplicando ou gerando conhecimento, uma relação de precedência existe como uma dependência natural entre conceitos nos processos cognitivos para este aprendizado (LAURENCE; MARGOLIS, 1999).

Neste contexto de conteúdos de aprendizagem, com necessidade e justificativa para organização da sequência de conteúdo para melhor orientação na aprendizagem, surgem alguns estudos com propósito de pesquisa científica para entendimento das relações de pré-requisito no contexto de aprendizagem e educação (BERGAN; JESKA, 1980), (OHLAND et al., 2004) e (VUONG et al., 2011).

Parte destas abordagens explora mais as relações léxicas entre itens (MILLER,1995) e relações de entidade refinadas em bases de conhecimento (MINTZ et al., 2009). Também há abordagens que tratam o problema como uma extração de relação ou como um problema de previsão de links usando a tradicional técnica de aprendizagem de máquina (TALUKDAR; COHEN, 2012) e (YANG et al., 2015).

Esforços recentes neste tema de pesquisa têm sido feitos para consideração do problema com foco em entender melhor as relações de pré-requisito sob a ótica da relação semântica na linguística computacional (LIANG et al., 2015). Nesta abordagem, Liang et al. (2015) propõem uma métrica baseada em uma teoria linguística.

A teoria dos quadros semânticos (FILLMORE, 2006) coloca que, para entender um conceito, é necessário entender todos os conceitos relacionados em seu quadro semântico. Para Fillmore (2006), um quadro semântico é uma estrutura coerente de conceitos relacionados, onde as relações têm a ver com a maneira como os conceitos ocorrem simultaneamente em situações do mundo real. O termo "quadro" da teoria se refere a qualquer sistema de conceitos relacionados de tal maneira que, para entender qualquer um destes conceitos, é necessário entender toda a estrutura em que o conceito se encaixa; quando um dos conceitos dessa estrutura é introduzido em um texto ou em

uma conversa, todos os outros conceitos são automaticamente disponibilizados.

Ou seja, Fillmore (2006) considera que o conhecimento do quadro é necessário para um conhecimento adequado das palavras referentes aos conceitos no quadro: uma palavra ativa o quadro, destaca conceitos individuais dentro do quadro e frequentemente determina uma certa perspectiva na qual o quadro é visualizado. Por exemplo, para entender a palavra vender, você precisa ter conhecimento sobre a situação da transferência comercial. Isso inclui, além do ato de vender, uma pessoa que vende, uma pessoa que compra, bens a serem vendidos, dinheiro ou outra forma de pagamento e assim por diante. No exemplo de transação comercial padrão, por exemplo, o conceito "venda" interpreta a situação da perspectiva do vendedor e "compra" da perspectiva do comprador.

Assim, a métrica de Liang et al. (2015), derivada da teoria dos quadros semânticos, mede as relações de pré-requisito baseada na simples observação do aprendizado humano, ou seja, se para aprender um conceito A é necessário se referir ao conceito B para muitos conceitos relacionados de A mas não vice-versa, então é mais provável que B seja um pré-requisito de A do que A seja pré-requisito de B. Ou seja, a métrica proposta por Liang et al. (2015), chamada distância de referência (RefD – *Reference Distance*), mede a extensão na qual o conceito *A* requer o conceito *B* como um pré-requisito A – calculando o quão diferentemente os dois conceitos se relacionam um a outro.

A métrica RefD é promissora por apresentar bons resultados, mas como Liang et al. (2015) afirmam, há oportunidades de melhoria e aprimoramentos na métrica através da exploração de diferentes representações semânticas e extração de relações léxicas entre os conceitos avaliados.

Nesta pesquisa, a métrica RefD foi modificada por meio do uso de estratégias de cortes na entrada do cálculo da fórmula, redução do espaço vetor e diferentes formas de

calcular a função de ponderação apresentada originalmente pelos autores. A nova métrica proposta nesta dissertação foi denominada OP-RefD – *Optimized Performance* RefD – e foi aplicada no grafo de conhecimento da DBpedia.

1.2. Problema

A identificação automática das relações de pré-requisitos entre conceitos é um tema de pesquisa que tem se ampliado em relação ao contexto de sua aplicação. Entre as áreas de sua utilização, a educacional demostra um importante potencial (CHANGUEL et al., 2015). Contudo, apresentando complexos desafios. O aumento da disponibilidade de recursos de aprendizado on-line criou novas oportunidades para alunos independentes, mas também aumentou a dificuldade de planejar um curso (SAYYADIHARIKANDEH et al., 2019).

A identificação manual dessas relações de pré-requisito entre os recursos ou conceitos de aprendizado é cara em termos de tempo (SAYYADIHARIKANDEH et al., 2019). Em um processo de conhecimento (trilha percorrida por meio de conceitos interligados de um determinado assunto), de aprendizagem autônoma, a identificação das relações de pré-requisitos entre conceitos é a base para um processo adequado de aprendizagem (CHANGUEL et al., 2015).

No atual processo educacional, a estruturação da sequência de conteúdo e da relação de pré-requisitos entre conceitos tem se desenvolvido por meio da intervenção de uma avaliação humana, de um especialista de domínio (como por exemplo, um professor da temática estudada). Entretanto, em um movimento de automatização deste processo de identificação das relações de pré-requisitos entre conceitos são propostas soluções, como a RefD (LIANG et al., 2015).

Na literatura científica, trabalhos recentes resolvem este problema de previsão automática de pré-requisito entre conceitos utilizando técnicas de aprendizagem de

máquina (LIANG et al., 2017), (MANRIQUE et al., 2018), (MANRIQUE et al., 2019), (ROY et al.,2019), (SAYYADIHARIKANDEH et al., 2019), (ZHOU; XIAO, 2019). Apesar de apresentarem alta eficácia nesta previsão, não se preocupam com o elevado tempo de processamento nas soluções, aspecto importante a ser considerado para soluções online.

Uma das poucas propostas que resolvem o problema com algoritmos de cálculos simples é o trabalho de Liang et al. (2015). A métrica RefD de Liang et al. (2015) apresentou melhora nas medidas estatísticas em termos de precisão, recall e medida F1 no estudo comparativo realizado pelos autores com outro método que utilizava aprendizagem de máquina.

Porém, os pesquisadores apontam a necessidade de extrapolar os contextos de aplicação, tanto em relação ao domínio dos conjuntos de dados avaliados, quanto na utilização de outras bases de conhecimentos com representações semânticas mais sofisticadas (além da Wikipedia, utilizada por eles). Também são apontadas possibilidades de novas propostas de variações na métrica, utilizando tanto medidas que avaliam relações semânticas entre os conceitos, quanto medidas que usam extrações automáticas de relações léxicas.

Logo, supõe-se que um método adaptado, OP-RefD, modificado da original, explorando um grafo de conhecimento que apresenta relações mais amplas entre seus conceitos ampliará a capacidade de identificar automaticamente (sem intervenção humana) as relações de pré-requisito entre conceitos. Além disto, supõe-se que estratégias de cortes com redução do espaço de entrada da fórmula da métrica otimizarão o tempo de processamento da RefD, criando um método com um desempenho melhor em termos de eficiência (tempo de resposta) nesta previsão.

Acredita-se que o desenvolvimento de uma métrica otimizada em termos de

desempenho (tempo de resposta), sem perda de eficácia (medidas estatísticas) é uma contribuição relevante para as áreas que necessitam de uma identificação imediata dessas relações em ambientes online.

1.2.1. Questão de Pesquisa

Para este trabalho, foi definida a seguinte questão de pesquisa – a qual estabelece o escopo do estudo experimental.

RQ1: O método adaptado proposto encontra resultados melhores (em termos de medida estatística F1) em menor tempo de execução quando comparado ao método original ao ser aplicado em uma estrutura de dados (grafo de conhecimento) mais amplo em termos de existência de mais relações estruturais entre conceitos?

Para a questão, considerando a fórmula do RefD, existem possibilidades de aplicação de diferentes funções de ponderação na medição da relação entre os conceitos ao se explorar as estruturas existentes no grafo. Outra possibilidade é o desenvolvimento de novas estratégias de corte das entradas usadas na fórmula do cálculo, uma vez que o grafo também traz uma quantidade maior de conceitos relacionados ao conceito em avaliação, dos quais apenas os mais relevantes ao conceito avaliado nos interessam.

A expectativa é que as estratégias mencionadas para o método adaptado consigam resolver o problema de previsão automática, com resultados melhores aos encontrados pelo método original, porém consumindo menos tempo de execução.

1.3. Objetivo

Este trabalho de pesquisa tem como objetivo desenvolver uma métrica modificando a RefD de Liang et al. (2015) para a identificação automática das relações de pré-requisito entre conceitos.

Para atingir este objetivo, desenvolver-se-ão estratégias de corte de entradas

(redução do espaço de entrada) e diferentes formas de cálculo do peso para o RefD. Para verificação dos resultados, serão utilizados *datasets* conhecidos, com relações de pré-requisito entre conceitos previamente determinado por humanos. O cálculo do RefD original também será submetido a testes usando o mesmo *dataset*, permitindo a comparação.

A intenção é comparar experimentalmente o novo método proposto com a métrica RefD de Liang et al. (2015) e avaliar os ganhos obtidos na previsão automática das relações de pré-requisito entre conceitos.

1.4. Método

Para o desenvolvimento desta pesquisa, a metodologia utilizada será uma abordagem experimental quantitativa. Na pesquisa quantitativa, os resultados podem ser quantificados, a qual está centrada na objetividade e recorre à linguagem matemática para descrever as causas de um fenômeno, as relações entre variáveis (FONSECA, 2002).

Seguindo um rigoroso planejamento, a pesquisa experimental inicia as etapas de pesquisa com a problemática e as hipóteses delimitando as variáveis do fenômeno estudado (TRIVIÑOS, 1987). Consiste em determinar um objeto de estudo, selecionar as variáveis que seriam capazes de influenciá-lo, definir as formas de controle e de observação dos efeitos que a variável produz no objeto (GIL, 2007). A pesquisa experimental seleciona grupos de assuntos relacionados, os quais são submetidos a distintos tratamentos, verificando as variáveis e checando se as discrepâncias observadas nas respostas são estatisticamente significantes (FONSECA, 2002).

Com o objetivo de validar ou descartar as hipóteses e avaliar a proposta desta pesquisa, foram realizados experimentos com a proposta de métrica – variações da métrica RefD (LIANG et al., 2015) em um grafo de conhecimento e a posterior análise quantitativa dos resultados obtidos. Nos experimentos, as variações ocorreram no

desenvolvimento de estratégias de cortes para redução das entradas no cálculo do RefD e em diferentes formas de cálculo da função de ponderação utilizada na métrica.

Por fim, avaliar o quanto as variações citadas acima alteram o resultado em um determinado *dataset* com relação de precedência de pares conceituais avaliados por humanos. Desta forma, espera-se obter resultados mais próximos ao de um *dataset* conhecido e com relações de pré-requisitos determinadas previamente por humanos.

1.5. Contribuições

Acredita-se que o resultado desta dissertação pode contribuir na resolução de problemas que necessitam previamente de uma identificação precisa da relação de pré-requisitos entre conceitos, como por exemplo a recuperação de objetos educacionais relevantes, bem como sugestões de sequência de aprendizagem.

Acredita-se que a contribuição seja relevante, principalmente ao se considerar ambientes online, os quais exigem um tempo de resposta imediato. Espera-se, então, contribuir com uma melhor performance em termos de tempo de execução na avaliação de relações de pré-requisito entre conceitos.

Assim, esta pesquisa analisa se há perda de eficácia (precisão dos resultados) e se há ganho de eficiência (tempo de processamento) quando um par conceitual é avaliado em relação a sua relação de precedência.

1.6. Organização da Dissertação

Esta dissertação está organizada em seis capítulos, sendo que o primeiro compreende a introdução e sua temática, abordando o problema de pesquisa, o objetivo, a metodologia e as contribuições da pesquisa.

Os demais capítulos do presente trabalho estão organizados da seguinte forma: o Capítulo 2 faz uma breve conceitualização dos principais fundamentos envolvidos, além

de apresentar alguns dos trabalhos relacionados; o Capítulo 3 descreve a solução proposta; no Capítulo 4 são apresentados os experimentos realizados. O capítulo 5 trata da análise dos resultados encontrados, enquanto no Capítulo 6 o trabalho é finalizado e são apresentadas as discussões das conclusões e perspectivas futuras.

2. FUNDAMENTAÇÃO E TRABALHOS RELACIONADOS

Neste capítulo são apresentadas a Wikipedia e a DBpedia, elementos fundamentais para realização da investigação das relações de pré-requisito entre conceitos explorando relações em um grafo de conhecimento. Em seguida, são apresentados os principais estudos relacionados ao tema desta dissertação e que serviram de base para sua construção.

2.1. A Construção do Conhecimento no Ciberespaço

No atual contexto tecnológico, o ciberespaço é um território que possibilita aos indivíduos desenvolverem uma rede de relacionamentos, independente das barreiras geográficas, econômicas, sociais, políticas e culturais (BICALHO; MORAIS, 2016). Um espaço virtual no qual trafegam um volume exponencial de dados, informações e conhecimentos.

O ciberespaço desterritorializa o conhecimento e promove uma cultura de coletividade, de compartilhamento, possibilitando aos indivíduos buscarem por informações e aprender novos conhecimentos em um espaço de inteligência coletiva (LÉVY, 2001). Contudo, o ciberespaço também fomenta a individualidade do indivíduo em meio a uma cultura coletiva (LÉVY, 2001).

No processo de aprendizagem, por exemplo, o indivíduo precisará buscar sua trilha de informações e conceitos para alcançar novos conhecimentos (CHANGUEL et al. 2015). Precisará buscar os conceitos que antecedem, que são pré-requisitos para a aprendizagem de outros para, então, alcançar seus objetivos de estudo.

Para Lévy (2001), a inteligência coletiva surge como uma sustentável forma de pensar por meio das conexões sociais viabilizadas pela rede de indivíduos no ciberespaço – na Internet. E neste ciberespaço há contextos de produção coletiva do conhecimento como a Wikipedia.

2.2. Wikipedia

A Wikipedia é caracterizada como uma enciclopédia, em plataforma livre e online na qual a todos os usuários – leitores – é permitido a atualização de seu conteúdo por meio da inserção e atualização de informações, expressos por artigos (LASLIE, 2003). Logo, a Wikipedia simboliza a moderna expressão da produção de conhecimento coletivo.

De fato, esta plataforma traduz a inteligência coletiva dos dias atuais, uma vez que é segmentada em diversos idiomas e cada um deles é suportado por um único subdomínio no domínio wikipedia.org. A Wikipedia, em sua última atualização de estatísticas de abril de 2020[1] conta com cerca de 309 idiomas ativos[2]. Em sua versão em inglês, são mais de 6 milhões de artigos e mais de 140 mil usuários ativos.

A Wikipedia faz parte do dia a dia de usuários da Web. Um internauta, ao navegar em uma ferramenta de busca procurando por algum conceito ou algum evento específico, frequentemente recebe, como uma das primeiras respostas, a página de um dos artigos da Wikipedia que exibe a definição do conceito buscado.

No entanto, nem sempre os usuários tem o entendimento do significado do conceito lendo todo o seu artigo da Wikipedia. Isso ocorre porque muitos não possuem conhecimento prévio para entendê-lo. Neste cenário, torna-se necessário aos mesmos a leitura prévia de algum conhecimento relacionado para ajuda-los a entender o conteúdo

[1] https://en.wikipedia.org/wiki/Wikipedia:Statistics
[2] https://en.wikipedia.org/wiki/Wikipedia#Language_editions

do conceito (ZHOU; XIAO, 2019). Em contrapartida, este trabalho de compreender um outro conceito relacionado pode ser facilitado pela própria estrutura da Wikipedia.

Isto acontece porque os artigos da enciclopédia são elaborados não apenas com texto livre, mas também utilizando recursos estruturados disponíveis. Um deles é a existência de vários links que conectam palavras/conceitos a outros artigos da Wikipedia.

Assim, seu alcance e sua estruturação se apresentam como uma potente ferramenta de conhecimento; e julgando pelo crescente número de artigos na comunidade científica, tem sido reconhecida como um recurso de escala e utilidade excepcionais (MEDELYAN et al., 2009). Em seu artigo sobre a utilização da Wikipedia através da extração de relações de pré-requisito entre seus conceitos, Zhou e Xiao (2019) apresentam diferentes frentes de trabalhos científicos nos quais os pesquisadores exploram a enciclopédia dentro deste tema.

Entre os destaques estão a utilização dos artigos da Wikipedia na determinação da ordem de aprendizado de materiais didáticos em cursos online através das relações de pré-requisito (LIMONGELLI et al., 2015), (DE MEDIO et al., 2017). A utilização dos conceitos para anotar objetos de aprendizagem e de suas relações no auxílio para identificação das relações de pré-requisitos entre os objetos de aprendizagem também aparecem em alguns trabalhos. Outro tema de destaque em trabalhos são aqueles que utilizam as relações de pré-requisitos conceituais da Wikipedia na criação de mapas conceituais em materiais de aprendizagem, como cursos universitários (YANG et at., 2015), (LIANG et al. 2017) e livros didáticos (WANG et al., 2016), (WANG; LIU, 2016) e MOOCs (PAN et al., 2017).

A abordagem proposta nesta dissertação é um método adaptado para identificação automática de pré-requisitos entre conceitos. A adaptação foi feita a partir da pesquisa apresentada por Liang et al. (2015), os quais focaram a aplicação de seu método utilizando

o vasto conhecimento coletivo disponível sobre conceitos na Wikipedia. Para a adaptação proposta, fez uso de um grafo de conhecimento criado a partir da estrutura da Wikipedia – a DBpedia.

2.3. DBpedia

A DBpedia é uma base de conhecimento construída a partir das informações estruturadas e multilíngues obtidas da Wikipedia e publicada nos padrões da *Linked Data* e *Semantic Web* (LEHMANN et. al., 2015). Segundo Mendes et al. (2011), metade das informações constantes na DBpedia são disponibilizadas no formato de ontologia interdomínios, com classes variadas e de diferentes granularidades.

A DBpedia tem como premissa o fornecimento de dados e a capacidade de realizar consultas – pesquisas no repositório de dados da Wikipedia. Através destas, são extraídos dados estruturados que podem ser utilizados pelos usuários para responder consultas expressivas. Estes dados, por sua vez, são extraídos de locais no domínio (e subdomínios) da Wikipedia, por meio de sentenças *Resources Description Framework* – RDF – um arcabouço para representar informações na web (BERNERS-LEE et al., 2001). Posteriormente, esses dados são expostos para consultas por meio de *endpoints* SPARQL *Query Language* para RDF.

O processo de construção da DBpedia classifica as entidades existentes nos artigos da Wikipedia de acordo com a DBpedia *Ontology*. Sendo uma ontologia multidomínio, a DBpedia *Ontology* (release 2016-10[3]), na versão em inglês, descreve 6,6 milhões de entidades e um total de 18 milhões de recursos na DBpedia – o que consiste em 13 bilhões de informações (triplas de RDF).

Na DBpedia *Ontology* os conceitos correspondem às categorias semânticas que

[3] https://wiki.dbpedia.org/downloads-2016-10

comumente são requisitadas em tarefas de *Named Entities Classification* – NEC em níveis superiores. Para Lehmann et al. (2015), a DBpedia é dividida em quatro categorias: *Mapping-Based Infobox Extraction*; *Raw Infobox Extraction*; *Feature Extraction*; *Statistical Extraction*.

A DBpedia, versão semântica da Wikipedia, foi escolhida para os experimentos por ser um grafo de conhecimento aberto, em constante evolução e estendido por uma comunidade global. Seu formato estruturado permite a extração de uma série de informações entre conceitos, tais como relacionamentos e todas as características de dados conectados abertos que podem ser explorados neste contexto de definição de pré-requisitos entre conceitos.

Além disso, sua capacidade e facilidade de extração de informações através da realização de consultas SPARQL nas sentenças RDF poderia impactar no desempenho do tempo de execução na avaliação de pré-requisitos entre conceitos.

2.4. Relações de Pré-Requisito entre Conceitos

Para Houaiss (2009), um requisito é condição para se alcançar determinado fim; logo, um pré-requisito, por associação, é uma pré-condição para alcançar este fim. Por outro lado, conceito é a compreensão que alguém tem de uma palavra, uma noção, uma concepção ou ainda, uma ideia sobre algo (HOUAISS, 2009).

Aplicado ao contexto da aprendizagem, um pré-requisito é uma exigência a ser cumprida, que se faz necessária para avançar para uma nova etapa, ou seja, um conceito que dependa de outro para sua compreensão tem este outro como pré-requisito. Para apreender o conceito C, por exemplo, é necessário passar primeiramente pelos conceitos A e B.

2.4.1. Distância de Referência (RefD)

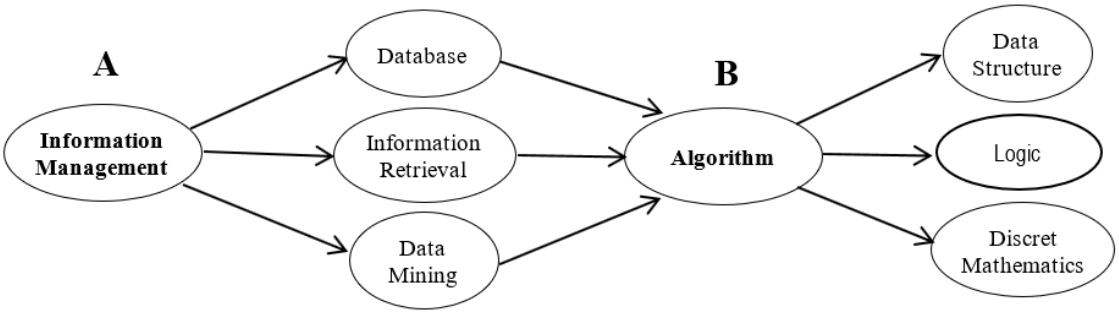
Desenvolvida por Liang et al. (2015), a métrica de distância de referência (RefD), tem como objetivo principal medir a extensão na qual o conceito A requer o conceito B como um pré-requisito. Esta métrica projeta uma função $f: C^2 \rightarrow R$ que mapeia o par de conceitos (A, B) para um valor real, onde C é o espaço conceitual. Assim, para definir como um conceito é representado no espaço conceitual C, foi utilizada a teoria de frame semântico (FILLMORE, 2006).

De acordo com a teoria, não é possível entender um conceito integralmente sem acessar todos os conhecimentos essenciais relacionados a ele. Tais conhecimentos podem ser vistos como um conjunto de conceitos relacionados. Assim, um conceito poderia ser representado por seus conceitos relacionados em C. Por exemplo, o conceito "*Deep Learning*" pode ser representado por conceitos como "*machine learning*", "*artificial neural network*", etc.

Uma relação mais comum e observável entre os conceitos é uma referência, que amplamente existe em várias formas, tais como hiperlinks, citações, notas etc. Embora uma única evidência de referência não indique uma relação de pré-requisito, um grande número dessas evidências pode fazer diferença (LIANG et al., 2015).

Por exemplo, se a maioria dos conceitos relacionados de A se referem a B, mas poucos conceitos relacionados de B se referem a A, então é mais provável que B seja um pré-requisito de A e não o contrário – como mostrado na Figura 1.

Figura 1: Estrutura de referência entre conceitos – B é pré-requisito de A



Fonte: Adaptado de (LIANG et al., 2015).

Para medir relações de pré-requisito, foi definida uma função de pré-requisito, denominada distância de referência - RefD:

$$RefD(A,B) = \frac{\sum_{i=1}^{k} r(c_i,B).w(c_i,A)}{\sum_{i=1}^{k} w(c_i,A)} - \frac{\sum_{i=1}^{k} r(c_i,A).w(c_i,B)}{\sum_{i=1}^{k} w(c_i,B)} \quad (2.1)$$

Onde:

- $C = \{c_1, c_2, \dots, c_k\}$ é o espaço conceitual
- $w(c_i, A)$ pondera a importância de c_i, para A
- $r(c_i, A)$ é um indicador de referência, que mostra se c_i refere-se a A, que podem ser links na Wikipédia, menções em livros, citações em artigos etc.
- K é o número total de conceitos no espaço conceitual

Assim, para colocar em prática o cálculo do $RefD$ é preciso:

- Especificar o espaço conceitual C;
- Especificar o peso w;
- Definir a função do indicador de referência r;

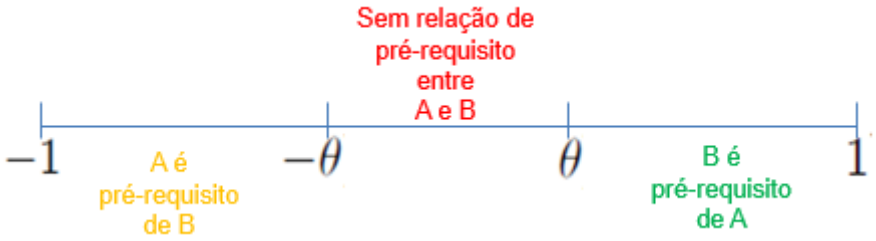
Os intervalos de valores possíveis calculados pela RefD são demonstrados na fórmula a seguir:

$$RefD(A,B) \in \begin{cases} (\theta, 1] & se\ B\ \text{é}\ pr\text{é} - requisito\ de\ A \\ [-\theta, \theta] & ,se\ não\ há\ relação\ de\ pr\text{é} - requisito\ entre\ A\ e\ B \\ (-1, -\theta], & se\ A\ \text{é}\ pr\text{é} - requisito\ de\ B \end{cases} \quad (2.2)$$

Onde θ é um *threshold* (valor limite) positivo, definido pelos autores como um ponto de corte para capturar as três possibilidades de relações de pré-requisito entre um

par conceitual. Para melhor ilustrar, a interpretação da função Distância de Referência para avaliação da relação de pré-requisito entre os dois conceitos é exibida na figura 2.

Figura 2: Interpretação do intervalo de valores calculados por RefD



Fonte: Adaptado de (LIANG et al., 2015).

2.5. Trabalhos Relacionados

No que se refere a temática desta pesquisa, há diversos trabalhos que aplicam a aprendizagem de máquina para identificar as relações de pré-requisitos entre os conceitos. São estas relações que permitem estruturar, de forma adequada, a trilha de aprendizagem de um aluno, em outras palavras, viabilizando uma possível automação. Os trabalhos aqui elencados fundamentam-se nessa premissa pré-requisitos entre conceitos.

Manrique et al. (2019) apresentam o uso de aprendizagem de máquina em seus experimentos, apontando que seu método supera o RefD (LIANG et al., 2015) aplicado aos mesmos *datasets*. Entretanto, a utilização de métodos de aprendizagem de máquina possui um grande obstáculo – o tempo de processamento –, que geralmente é superior a outros métodos. Em um processo de *e-learning*, por exemplo, o tempo de processamento para a estruturação da trilha de conhecimento é crucial. Tendo este aspecto em foco, nesta pesquisa optou-se por otimizar o método RefD (LIANG et al., 2015) para obter melhor desempenho na identificação das relações de pré-requisitos entre conceitos.

Além disto, a escolha em adaptar o método RefD de Liang et al. (2015) – ao invés do método proposto por Manrique et al. (2019) –, se deve ao fato do RefD se basear em um cálculo simples, com o potencial de apresentar melhores tempos de processamento em relação aos métodos de aprendizagem de máquina.

Talukdar e Cohen (2012) propõem uma abordagem que possibilita aos leitores, alunos, a compreensão de materiais técnicos complexos por meio de métodos estatísticos. Estes métodos modelam uma estrutura de pré-requisitos de um corpus, ou seja, o impacto semântico que os documentos terão no estado de conhecimento do indivíduo (aluno). Para isto, os pesquisadores utilizaram aprendizagem de máquina para prever a estrutura de pré-requisitos não apenas em relação ao texto de um documento, mas também aos dados gerados por *crowdsourcing*, como hiperlinks e logs de edição. Os experimentos utilizaram a Wikipedia como corpus e sugerem que, embora não seja óbvio que essa tarefa seja viável, existem *features* relativamente confiáveis para prever a estrutura de pré-requisitos usando métodos padrão de aprendizado de máquina. Os autores demonstram grande potencial em seu método e novas possibilidades de expansão da pesquisa.

Para Changuel et al. (2015), o objetivo de qualquer sistema de tutoria é o de fornecer recursos que se adaptem ao indivíduo/aluno em relação ao seu conhecimento prévio. Em seu trabalho, apontam o desafio de fornecer ao aluno uma trilha de documentos adequada, de forma que cada novo documento se baseie em conceitos já definidos nos documentos anteriores.

Assim, Changuel et al. (2015) descrevem um método de anotação de conceito que se baseia em técnicas de aprendizagem de máquina para prever a classe de cada conceito - pré-requisito ou resultado - com base em informações contextuais e recursos locais. A partir da categorização, os autores estabelecem um sequenciamento automático de recursos baseado na relação de precedência entre os recursos de aprendizagem.

Os resultados apontados pelos pesquisadores mostram consistência do sequenciamento proposto em relação à tabela verdade fornecida pelo autor do conteúdo online (CHANGUEL et al., 2015). O trabalho apresenta importantes conceitos da construção do conhecimento. Contudo, sua posposta não considera a utilização em

ambientes online que exigem tempo de resposta imediato.

Gasparetti et al. (2015) apresentam uma abordagem para auxiliar os professores na definição das relações de pré-requisito entre Objetos de Aprendizagem (*Learning Objects* – LO). Para tanto, utilizam técnicas de aprendizagem de máquina de análise semântica que identificam na Wikipedia conceitos relevantes em cada LO. Gasparetti et.al. (2015) escolhem um conjunto de características para explorarem o conteúdo e a estrutura da Wikipedia. Os resultados apresentados pelos autores mostram boa precisão quando são utilizados os conteúdos em texto da Wikipedia, mas baixa precisão quando são utilizadas características estruturais da Wikipedia.

Ao fim, os pesquisadores sugerem que o uso de projetos de dados conectados abertos, tais como a DBpedia (LEHMANN et al., 2015), Freebase (BOLLACKER et al., 2008) ou Yago (HOFFART et al., 2013) poderia enriquecer seus experimentos com resultados mais precisos, uma vez que proporcionam consultas a relacionamentos semânticos e propriedades associadas aos artigos da Wikipedia. Uma outra vantagem, segundo os autores, seria o acesso às informações relevantes de maneira mais eficaz (GASPARETTI et al., 2015). Assim como outras pesquisas, também não consideram a demanda no tempo de processamento que a aprendizagem de máquina exige para retornar um alto grau de precisão.

Liang et al. (2015) apontam que uma relação de pré-requisito tem a finalidade de descrever uma relação entre conceitos em cognição, educação e em diversas áreas. Contudo, na linguística computacional as relações semânticas carecem ser melhor estudadas. Logo, os pesquisadores investigam o problema de mesurar as relações de pré-requisito entre conceitos e, consequentemente, propõem uma métrica simples baseada em link, distância de referência – RefD. Esta métrica modela a relação medindo a diferença entre dois conceitos. Para sua validação, dois conjuntos de dados (sete domínios)

possibilitaram demostrar a singularidade da métrica, uma vez que o método baseado em desempenho superou os métodos existentes baseados em *machine learning*, aprendizado supervisionado. Os testes foram realizados, baseados na Wikipédia por meio de duas estratégias de ponderação: EQUAL e TFIDF. Por fim, os autores sugerem a aplicação do RefD em outros contextos e sugerem a mensuração das relações de pré-requisito ou ordens de leitura entre livros didáticos, além de sugerir ainda a possibilidade de incorporação do RefD nos modelos de sistemas supervisionados para aumentar a acurácia.

Liang et al. (2017) investigam meios para recuperar relações de pré-requisito (relações de dependência) dos cursos nas universidades que participaram do experimento. O modelo proposto infere os pré-requisitos do curso através da construção de um grafo conceitual das dependências observadas do curso, em vez de extrapolar os pré-requisitos para pares de conceitos dos cursos, como no trabalho anterior (LIANG et al., 2015). Para avaliar o problema de pesquisa, os pesquisadores criaram um conjunto de dados real para estudar empiricamente esse problema. Este conjunto é formado por uma lista de cursos de Ciência da Computação de 11 universidades dos Estados Unidos e de seus pares de conceito com rótulos de pré-requisito.

A diferença para o trabalho anterior (LIANG et al., 2015), é que o novo método proposto em 2017 é baseado em aprendizagem de máquina não supervisionada. Este é o ponto chave que justifica a escolha desta dissertação em explorar soluções baseadas na primeira proposta de Liang et al. (2015) para resolução do problema. O interesse na métrica de Liang et al. (2015) se dá justamente por ser um cálculo com melhor potencial a ser utilizado em ambientes online, uma vez que, por se tratar de um cálculo computacionalmente simples, não tem o alto custo de tempo de processamento como em soluções de aprendizagem de máquina.

Manrique et al. (2018) desenvolveram uma pesquisa acerca das estratégias automáticas que poderiam ser utilizadas para estabelecer relações de precedência entre recursos de aprendizagem. A análise apresentada verificou dois recursos de aprendizado analisando pré-requisitos entre os conceitos abordados pelos recursos de aprendizagem para estimar precedência da relação. Os resultados experimentais demonstraram a possibilidade de identificar a relação de precedência entre os recursos de aprendizagem por meio da identificação automática de relações de pré-requisito entre conceitos. O artigo foca em métodos de aprendizado de máquina para medir a relação de pré-requisitos entre conceitos.

Em outro trabalho, Manrique et al. (2019) abordam sobre as especificidades do problema da identificação de pré-requisitos conceituais para fornecer corretamente o conhecimento básico necessário para compreensão de um conceito particular. Para tanto, os pesquisadores exploram o grafo de conhecimento da DBpedia e suas relações semânticas com a finalidade de encontrar conceitos candidatos para um conceito-alvo. A abordagem utiliza algoritmos de aprendizado supervisionado que avalia e gera uma lista de pré-requisitos para o conceito de destino. Na validação, a precisão final obteve resultados promissores, os quais variaram entre 83% e 92,9%, principalmente variando em relação a configuração do parâmetro do método de corte.

Contudo, por mais que os resultados em ambos os artigos de Manrique et al. (2018 e 2019) sejam promissores, as propostas de solução para a automação da aprendizagem utilizando algoritmos de aprendizagem de máquina são custosas em termos de tempo de processamento. Este tipo de solução inviabiliza sua aplicação em ambientes online, pois a construção dinâmica e autônoma da trilha de conhecimento do aluno exigiria um maior poder de processamento e demandaria tempo elevado.

Sayyadiharikandeh et al. (2019) apontam sobre a grande disponibilidade de

recursos de aprendizado em plataforma *online* e a oportunidade de independência por parte do aluno na busca por conhecimento. Os pesquisadores propõem um método para identificação de relações de pré-requisito baseado em dados que ocorrem naturalmente - os padrões de navegação dos usuários na Wikipedia. Utilizam uma abordagem de aprendizado supervisionado com o treinamento de classificadores, os quais possibilitam perceber que a estrutura de rede de navegação pode ser usada para identificar dependências entre conceitos em vários domínios.

Roy et al. (2019) apresentam um método de aprendizado supervisionado (PREREQ) para inferir as relações de pré-requisito de conceito. O método foi projetado utilizando representações latentes de conceitos obtidos do modelo *Pairwise Latent Dirichlet Allocation* e de uma rede neural baseada em *Siamese Network Architecture*, o qual permite aprender pré-requisitos de conceito desconhecido dos pré-requisitos do curso e pré-requisito do conceito já rotulado. Segundo os pesquisadores, PREREQ supera as abordagens de ponta em conjuntos de dados de referência e pode aprender efetivamente com muito menos dados de treinamento, além de poder utilizar lista de reprodução de vídeos não rotuladas para aprender.

Zhou e Xiao (2019) apresentam um método para identificação de relações de pré-requisitos entre conceitos com experimentos em oito conjuntos de dados (inglês e chinês) da Wikipedia. Para desenvolver seu método, os pesquisadores criaram quatro grupos de recursos (baseados em link, categorias, conteúdo e tempo) com a finalidade de prever as relações de pré-requisito entre os pares de conceito. Posteriormente, coletaram os conceitos de ambos *datasets* (inglês e chinês) da Wikipedia e construíram oito conjuntos de dados (para treinamento e teste) composto de centenas de pares de conceitos. Segundo os pesquisadores seu método supera os métodos existentes de aprendizagem de pré-requisito.

Nestes últimos três artigos, de Sayyadiharikandeh et al. (2019), Roy et al. (2019) e Zhou e Xiao (2019), também são apresentados resultados promissores para o problema. Contudo, diferentemente da abordagem desta pesquisa, nenhum deles considera nos estudos a questão do tempo de processamento necessário para a utilização destes métodos. Como mencionado, o tempo de processamento é crucial em ambientes online (ao se considerar, por exemplo, ferramentas de busca ou repositórios de objetos de aprendizagem, onde a disponibilidade de conceitos tem uma escala maior, com projeção de crescimento exponencial).

Ao se pensar em *e-learning*, na construção de conhecimento por meio de plataforma digital, o fator tempo/espera de processamento e custo de processamento devem ser levados em consideração. Nos trabalhos anteriormente mencionados, com exceção do trabalho apresentado por em Liang et al. (2015), essa premissa é desconsiderada.

3. DISTÂNCIA DE REFERÊNCIA MODIFICADO – OP-RefD

Este capítulo tem o objetivo de descrever como foi a condução da solução proposta desenvolvida para o problema apresentado e para testar a hipótese desta dissertação. A seção 3.1 descreve sucintamente a forma como a métrica genérica RefD foi criada por Liang et al. (2015) e especializada no contexto da Wikipedia, enquanto as demais seções tratam de como a métrica foi utilizada no contexto da DBpedia e as modificações realizadas neste trabalho com intenção de explorar as relações entre conceitos existentes no grafo da DBpedia.

3.1. Implementação do RefD Original Usando a Wikipedia

Como uma enciclopédia aberta, amplamente utilizada, a Wikipedia fornece conhecimento relativamente atualizado e de alta qualidade (GABRILOVICH; MARKOVITCH, 2007). Além disso, os hiperlinks criados pelos editores da Wiki fornecem uma maneira natural de calcular a função do indicador de referência. Sendo assim, quando Liang et al. (2015) propuseram a métrica geral RefD, a especificaram utilizando a Wikipedia para testar sua validade.

Logo, a implementação foi realizada considerando o espaço conceitual C consistindo em todos os artigos da Wikipédia. O indicador de referência $r(c, A)$ representa se existe um link do artigo c da Wiki para o conceito A. Para o cálculo da função de peso $w(c, A)$ foram experimentados dois métodos de cálculo:

- EQUAL – o peso de um conceito c para o conceito A é representado pelos conceitos conectados a ele $L(A)$ com pesos iguais. Peso 1, se há ligação e peso 0, se não há.

$$w(c,A) = \begin{cases} 1 & c \in L(A) \\ 0 & c \notin L(A) \end{cases} \quad (3.1)$$

- TFIDF - o peso de um conceito c para o conceito A é representado pelos conceitos conectados a ele $L(A)$ com pesos TFIDF:

$$w(c,A) = \begin{cases} tf(c,A) * log\dfrac{N}{df(c)} & c \in L(A) \\ 0 & c \notin L(A) \end{cases} \quad (3.2)$$

Onde:

- $tf(c,A)$: número de vezes que c está relacionado a A;
- N: número total de artigos da Wikipedia;
- $df(c)$: número de artigos da Wikipedia em que c aparece;

Dentre os dois métodos de peso utilizados pelos autores, o TF-IDF apresentou melhores resultados de métricas estatísticas nas avaliações realizadas.

3.2. Implementação do OP-RefD Usando a DBpedia

A estratégia adotada foi testar diferentes funções de corte – ou seja –, formas de diminuir o número de conceitos do espaço conceitual C, os quais são a entrada para o cálculo da distância de referência, assim como proposto em (MANRIQUE et al., 2019). É importante citar que, para o desenvolvimento da experimentação desta pesquisa, foram necessárias algumas adaptações no cálculo da métrica RefD para sua utilização na DBpedia.

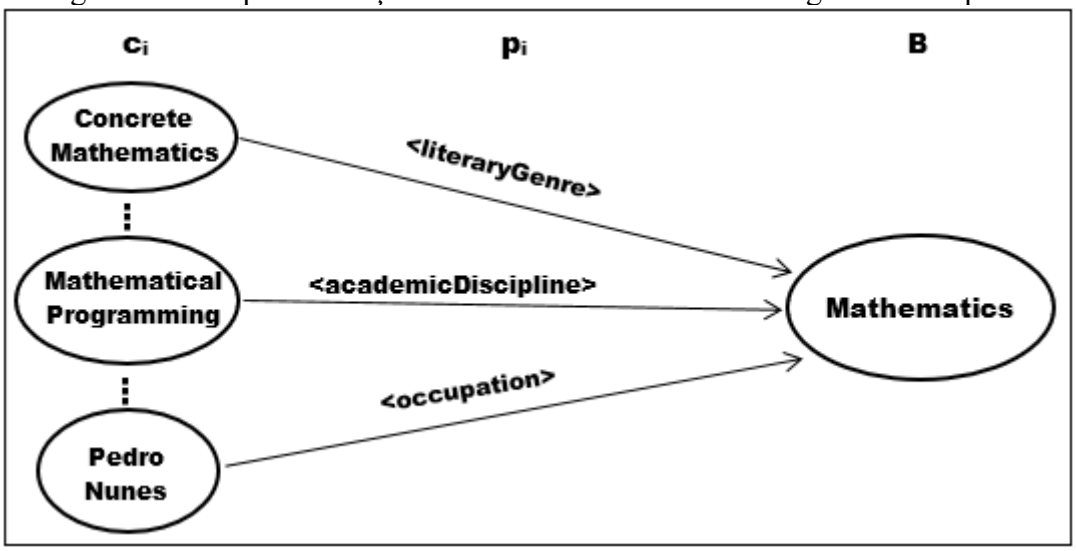
Para adaptação do RefD, a implementação proposta do OP-RefD leva em consideração que a função de ponderação $w(c,A)$ tem como entrada os conceitos vizinhos comuns no grafo da DBpedia, enquanto a função indicador de referência $r(c,A)$ representa se existe um caminho (uma tripla) de propriedade entre o destino e conceitos

conectados a ele.

Desta forma, para a função do indicador de referência, $r(c_i, B)$, considerando o conceito B avaliado, são recuperados todos os conceitos c_i distintos que apontam para B através de uma propriedade p_i. A figura 3 ilustra um exemplo para melhor entendimento.

A figura 3 também pode ser usada para justificar a ideia por trás da função de corte. Neste exemplo vemos alguns conceitos c_i que são considerados no cálculo do OP-RefD. Intuitivamente, percebe-se que o conceito "Pedro Nunes" tem muito menos relevância, semanticamente, para a avaliação dos conceitos que são pré-requisito do conceito avaliado "*Mathematics*", uma vez que indica se tratar de um nome próprio relacionado ao artigo, e não de um conceito em si. Analisando exemplos reais extraídos da estrutura do grafo da DBPedia, verifica-se que aqueles conceitos vizinhos considerados na fórmula que não tem muita relação semântica podem contribuir negativamente na assertividade do resultado do cálculo.

Figura 3: Exemplo da função de indicador de referência no grafo da DBpedia



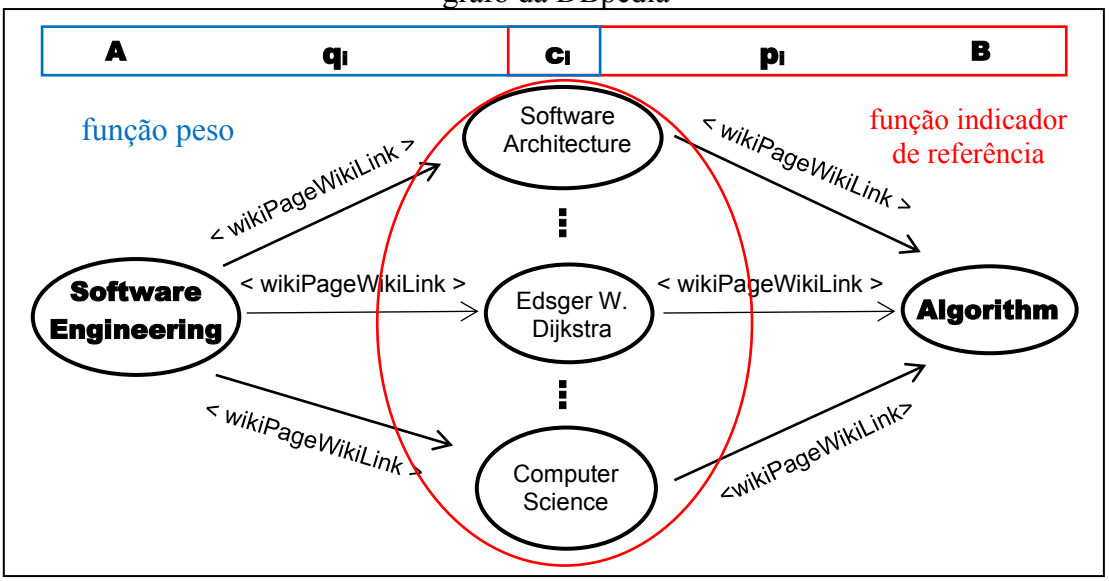
Fonte: Própria.

Desta forma, acredita-se que a estratégia de corte, com diminuição do espaço conceitual de entrada, seja adequada para aumentar a assertividade na identificação automática de pré-requisitos entre conceitos, pois eliminariam os conceitos que não tem

muita relação com o conceito em avaliação.

Para a função de peso, $w(c_i, A)$, considerando o conceito A avaliado, são recuperados todos os conceitos c_i relacionados a A através de uma propriedade q_i. Na prática, a parte do numerador dentro da fórmula (2.1) do RefD (função peso multiplicada por função de indicador) se constitui dos conceitos c_i para os quais A aponta através de uma propriedade q_i, e que sejam comuns com os conceitos c_i relacionados a B através de uma propriedade p_i, ou seja, conceitos vizinhos em comum entre A e B. A figura 4 ilustra um exemplo para melhor entendimento. A função de corte visa diminuir o tamanho desse espaço conceitual – a quantidade de conceitos entre A e B – conforme delimitado pela elipse vermelha indicada na figura 4.

Figura 4: Exemplo do relacionamento entre a função indicador e função de peso no grafo da DBpedia



Fonte: Própria.

Para melhor ilustrar as especificidades da aplicação do RefD nas duas diferentes bases de dados, a tabela 1 traça um paralelo do que representa cada elemento nas implementações da distância de referência na Wikipedia *versus* na DBpedia.

Tabela 1: Comparativo dos elementos do RefD na Wikipedia *versus* DBpedia.

Elemento do RefD	**Wikipedia**	**DBpedia**
$\boldsymbol{C}$: espaço Conceitual	todos os artigos da Wikipedia	todos os recursos da DBpedia (http://dbpedia.org/resource/)
$\boldsymbol{r(c_i, A)}$**:** função indicadora de referência diz se c_i refere-se a A	hiperlinks avançados criados pelos editores da Wiki diz se existe um link entre o artigo c_i para o artigo A	arestas do grafo da DBpedia diz se existe ligação no grafo da DBpedia de alguma propriedade do recurso c_i para A
$\boldsymbol{w(c_i, A)}$**:** função peso diz a importância de c_i para A	Implementados pesos EQUAL e TF-IDF	Implementados pesos EQUAL, TF-IDF, e outros discutidos nas próximas subseções

Fonte: Própria

As adaptações realizadas, necessárias para a experimentação, basearam-se nas características da DBpedia e de sua estrutura de grafo de conhecimento.

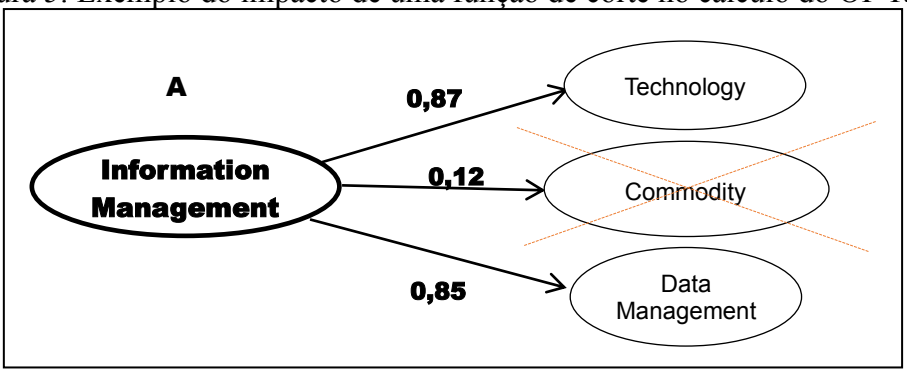
3.2.1. Adaptações no Cálculo do OP-RefD usando a Função de Corte

A função de corte visa diminuir os conceitos c_i comuns entre a função indicador de referência do conceito B, $r(c_i, B)$, e a função de peso com o conceito A, $w(c_i, A)$, das partes dos numeradores da fração da fórmula (2.1) do cálculo do RefD. Para tanto, considera apenas aqueles c_i comuns que, quando comparados com o conceito A usando alguma função de similaridade entre conceitos, estão acima de um *threshold*, um limiar mínimo de semelhança.

Este limiar é determinado experimentalmente, de forma a maximizar o valor da medida F1 (mais detalhada na seção 5.1), entre cada um dos conceitos c_i e A (numerador da primeira parte da fórmula), e semelhança entre cada um dos conceitos c_i e B (numerador da segunda parte da fórmula). A figura 5 exemplifica uma situação de corte,

ilustrando os conceitos entre o par conceitual *Information Management* (A) e *Artificial Intelligence* (B) – este ultimo suprimido da imagem.

Figura 5: Exemplo do impacto de uma função de corte no cálculo do OP-RefD



Fonte: Própria.

Durante a avaliação de pré-requisito entre o par conceitual mencionado, o esquema exibe a estrutura da função de peso aplicada ao conceito *Information Management*, destacando o valor do peso de cada aresta dos conceitos para os quais este conceito A aponta. O conceito *Commodity*, antes considerado no cálculo da função de peso, como um dos conceitos c_i para os quais o conceito *Information Management* aponta, quando aplicada uma função de corte, não foi computado no cálculo do OP-RefD. Assim, o módulo desenvolvido para avaliação da função de corte, retirou os c_i do espaço conceitual C' considerado para o OP-RefD modificado, calculado após corte.

A seguir, serão descritos os diferentes métodos de cálculo testados tanto como método de função peso quanto como método de função de corte para diminuição do espaço conceitual dos conceitos em comum considerados no RefD.

3.2.2. Frequência TF-IDF da Propriedade

Analogamente à função de peso TF-IDF (abreviação do inglês *term frequency-inverse document frequency*) adotada e contextualizada por Liang et al. (2015), foi pensada uma outra função baseada em frequência, porém ao invés de considerar a frequência dos conceitos em comum, considerou-se a frequência da propriedade que

conecta os conceitos.

Originalmente o TF-IDF é uma medida estatística que indica a importância de uma palavra em um documento em relação à uma coleção de documentos ou corpo linguístico. A ideia por trás desta medida é que os termos mais frequentes em um documento são mais importantes, pois podem ser maiores indicativos do tópico do documento. Esta métrica é muito utilizada como uma medida de ponderação em aplicações de recuperação de informações e mineração de dados.

Assim, foi desenvolvida a função $TF - IDF_{prop}$ para medir a importância da propriedade (aresta no grafo DBpedia) que liga os conceitos c_i conectados entre A e B – conceitos avaliados.

$$TF - IDF_{prop}(p, A) = \begin{cases} tf(p, A) * log \dfrac{N}{df(p)} & p \in L(A) \\ 0 & p \notin L(A) \end{cases}$$

Onde:

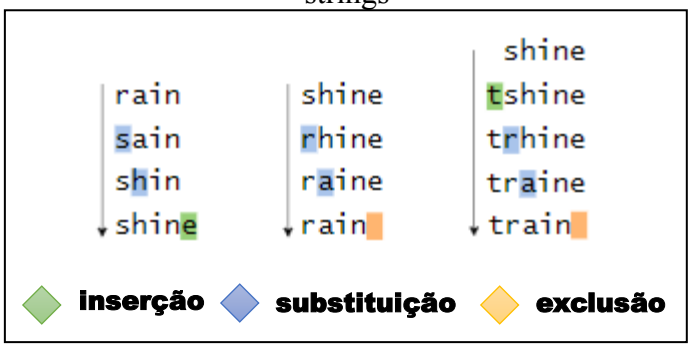
- $tf(p, A)$: número de vezes que p está relacionado ao conceito A;
- N: número total de propriedades (arestas) no grafo da DBpedia;
- $df(p)$: número de conceitos da DBpedia em que p aparece;

3.2.3. Levenshtein – Distância Mínima de Edição

Nas áreas de teoria da informação, linguística e ciências da computação, a distância de Levenshtein é uma métrica usada para calcular a distância de edição entre duas *strings* (sequência de caracteres). Sendo assim, a distância de Levenshtein (LEVENSHTEIN, 1966) pode ser usada para medir a similaridade entre duas palavras e consiste no número mínimo de edições de um caractere (inserções, exclusões ou substituições) necessárias para transformar uma palavra em outra, conforme demonstrado na figura 6. Desta forma, quanto maior a distância de Levenshtein, menor a similaridade

entre duas palavras.

Figura 6: Exemplo de operações possíveis no cálculo da distância de edição entre duas strings



Fonte: Baseado na Devopedia 2019 [4].

A distância de Levenshtein $LD(A, B)$ entre dois conceitos A e B tem um valor máximo igual ao comprimento do conceito com maior número de caracteres entre os dois conceitos entrada, ou seja, não pode ficar pior que isso. Portanto, para comparação da distância de edição, foi considerado um índice de similaridade normalizado, onde valores próximos a 0 indicam que os conceitos não são muito similares e valor igual a 1 é para palavras iguais. Assim, a distância normalizada $LD_{norm}(A, B)$ entre dois conceitos A e B pode ser definida como:

$$LD_{norm}(A, B) = 1 - \frac{LD(A, B)}{\max\left[len(A), len(B)\right]}$$

Onde,

- $len(A)$: número de caracteres que formam o conceito A;
- $\max\left[len(A), len(B)\right]$: maior valor entre $len(A)$ e $len(B)$;

3.2.4. Similaridade Semântica – Word2VEC

A similaridade semântica entre palavras/termos, sentenças ou documentos, é uma métrica definida sobre um conjunto de termos ou documentos que mede o quanto um termo é semelhante a outro em relação ao seu significado ou conteúdo semântico. Neste

[4] https://devopedia.org/levenshtein-distance

trabalho, para avaliar o quanto um conceito é similar a outro, como função de peso ou corte no cálculo do RefD, foi utilizada a biblioteca spaCy[5].

O spaCy é uma biblioteca de software de código aberto para processamento avançado de linguagem natural, escrita nas linguagens de programação Python e Cython e, dentre outros métodos, possui a capacidade de realizar avaliação de similaridade semântica entre termos. Diversos trabalhos científicos recentes comprovam a eficiência desta biblioteca para esta finalidade (PAWAR, 2018), (GAO et al., 2019), (ZHU; IGLESIAS, 2018), (VAKARE et al., 2019), (LALA et al., 2019), (SCHMITT et al., 2019), o que justifica a escolha deste caminho como uma possibilidade de maximizar os resultados da medida modificada de distância de referência entre pares de conceitos.

A biblioteca utilizada carrega um modelo de vetores de palavras gerado usando o algoritmo Word2Vec (MIKOLOV et al., 2013). Resumidamente, o Word2Vec é um grupo de modelos relacionados que são usados para produzir *word embeddings* (incorporação de palavras), em que termos do vocabulário são mapeadas para vetores de números reais. Conceitualmente, envolve uma incorporação matemática de um espaço com muitas dimensões por palavra para um espaço vetorial contínuo com uma dimensão muito menor.

Foi utilizado o modelo spaCy *en_core_web_lg* (HONNIBAL; MONTANI, 2017). Este modelo spaCy pré-treinado fornece vetores GloVe de 300 dimensões com cerca de um milhão de palavras no idioma inglês[6].

O método utilizado da spaCy faz o cálculo da similaridade semântica entre *tokens* como um valor entre 0 e 1 – calculado através da similaridade do cosseno entre os vetores gerados para cada conceito. Para dois conceitos comparados, quanto mais próximo de 1 for a pontuação semântica retornada, mais semelhante é um conceito do outro. Quanto mais próximo de 0, mais distinto é um conceito do outro.

[5] https://spacy.io/
[6] https://spacy.io/usage/vectors-similarity

4. AVALIAÇÃO DO OPTIMIZED PERFORMANCE RefD (OP-RefD)

Este capítulo tem o objetivo de descrever o estudo experimental com o intuito de avaliar os diferentes métodos propostos como variações para funções de peso e de corte e aplicação no cálculo da distância de referência. Os métodos de peso gerais criados por Liang et al. (2015) foram também implementados utilizando a DBpedia para serem comparados às funções diferenciadas criadas.

4.1. *Datasets* Avaliados

Na investigação da questão de pesquisa, para verificar a influência das funções de peso e de corte no cálculo do RefD e testar a validade das diferentes versões do OP-RefD, foram utilizados dois conjuntos de dados de pares de conceitos – *datasets* com a relação de precedência já definidas.

Para avaliação do impacto das funções descritas no capítulo 4, os dois *datasets* utilizados foram os mesmos aplicados por Liang et al. (2015) em sua abordagem para implementação do RefD original: O *dataset* RefD2015[7] (LIANG et al., 2015) e o *dataset* UCD[8] – *University Course Dependencies* (LIANG et al., 2017). Posteriormente, Manrique et al. (2018) fazem uso destes *datasets* para seus experimentos, entretanto, com a aplicação de técnicas de aprendizagem de máquina.

O *dataset* RefD2015 é um conjunto de dados composto por pré-requisitos relacionados a cursos. Foi construído com a ajuda de informações disponíveis no site do

[7] *Dataset* 1 <https://github.com/harrylclc/RefD-dataset>
[8] *Dataset* 2 < https://github.com/harrylclc/eaai17-cpr-recover>

curso de uma universidade contendo relações de pré-requisito entre os cursos.

Os pares conceituais de pré-requisitos foram obtidos rastreando o site com uma ferramenta *web scraper*. Em seguida, os cursos foram vinculados aos artigos da Wikipedia usando regras simples, como correspondência de título e similaridade de conteúdo. Todos os pares conceituais foram verificados e ajustados manualmente por dois especialistas de domínio para remoção de pares com rótulos incorretos.

Para obtenção de exemplos negativos na relação entre os conceitos, foram amostrados aleatoriamente 600 pares usando conceitos que aparecem nos pares de pré-requisito. Este primeiro *dataset* consiste em dois domínios: Math e CS (*Computer Science*). A tabela 2 exibe as estatísticas do *dataset* RefD2015 exibindo o quantitativo de pares conceituais avaliados e a quantidade de pares conceituais com relação de pré-requisito encontrados entre eles.

Tabela 2: Estatísticas do *Dataset* RefD2015

Dataset	**Domínio**	**# Pares**	**# Pré-requisitos**
RefD2015	CS	678	108
	MATH	658	75

Fonte: Baseado em Liang et al. (2015).

O *dataset* UCD – *University Course Dependencies* – é um conjunto de dados composto de pré-requisitos criado por Liang et al. (2017) com base em dados coletados de onze universidades dos Estados Unidos, de cursos com foco em ciência da computação ou departamentos semelhantes a ciência da computação.

Foi desenvolvido através de uma ferramenta *web scraper* executada para extrair as dependências de disciplinas de curso dos catálogos on-line destas universidades. Após a coleta das informações de descrição do curso, foram aplicadas ferramentas automáticas para extração de conceitos da Wikipedia. O *dataset* UCD é composto por um domínio

único – *Computer Science* – e suas estatísticas, conforme tabela 3, são exibidas a seguir.

Tabela 3: Estatísticas do *Dataset* UCD

Dataset	**Domínio**	**# Pares**	**# Pré-requisitos**
UCD	CS	1685	1004

Fonte: Baseado em Liang et al. (2017).

Para exemplificar, a tabela 4 exibe alguns exemplos de um dos *datasets* avaliados experimentalmente nesta pesquisa. Eles são os conceitos A e B que servem de entrada para o cálculo do OP-RefD, usando a estrutura da DBpedia.

Os pares conceituais da tabela 4 mostra uma extração dos pares conceituais que foram avaliados por especialistas de domínio, lembrando que esta foi considerada a base para comparações feitas nesta dissertação. Ou seja, os valores verdadeiros contra os quais serão confrontados os valores previstos após cálculo do OP-RefD, seguida de avaliação de acordo com limiar definido.

A interpretação da tabela é a seguinte: quando o conceito B foi considerado pelo especialista um pré-requisito de aprendizado para o conceito A, a coluna "Pré-requisito" foi preenchida com o valor 1. Quando o conceito B não foi considerado pelo especialista um pré-requisito de aprendizado para conceito A, a coluna "Pré-requisito" foi preenchida com valor 0 (se a relação for que o conceito A é pré-requisito do conceito B, o valor também é 0).

Tabela 4: Pares conceituais com indicativo de relação de pré-requisito, conforme avaliação de especialistas de domínio

id	**Conceito A**	**Conceito B**	**Pré-Requisito**
1	Network_security	Computer_network	1
2	Network_security	Algorithm	1
3	Software_engineering	Algorithm	1

4	Distributed_object	Computer_graphics	0
5	Theory_of_computation	Object-oriented_programming	0
6	Software_project_management	Translator_(computing)	0

Fonte: Própria.

4.2. Ambiente Computacional

Os testes conduzidos durante a implementação dos algoritmos de avaliação de pré-requisitos entre conceitos foram executados em um servidor na plataforma de computação em nuvem da AWS[9] – *Amazon Web Services* – utilizando sistema operacional Linux Ubuntu Server 16.04, SSD Volume Type, com uma instância do tipo t2.2xlarge (Processadores Intel Xeon de alta frequência, 8 núcleos) com 32GB de memória RAM.

Adicionalmente, é válido citar que o ambiente necessário para execução dos experimentos apresentados neste capítulo, exigiu a instalação de um servidor Virtuoso *Universal Server - Open Source Edition*[10], que provê serviços tanto de servidor de aplicação como de gerenciamento de banco de dados para triplas RDFs, fornecendo também serviço de consultas SPARQL.

Desta forma foram carregados no virtuoso os *dumps* da DBpedia 2016-10[11]. As triplas referentes ao idioma inglês foram carregadas, uma vez que os *datasets* utilizados trazem pares conceituais avaliados neste idioma. Além disto, o código implementado foi desenvolvido utilizando a linguagem de programação Python[12], em sua versão 3.0.

[9] https://aws.amazon.com/
[10] http://vos.openlinksw.com/owiki/wiki/VOS/
[11] https://wiki.dbpedia.org/downloads-2016-10/
[12] https://python.org/

4.3. Configuração dos Experimentos

Foi utilizada a DBpedia para realização do cálculo da métrica proposta (OP-RefD) para previsão de relações de pré-requisito e avaliação de sua assertividade. Ou seja, para avaliar automaticamente a relação de precedência de um par conceitual dos conjuntos de dados avaliados, é utilizada a correspondência na DBpedia de cada um dos conceitos do par. Para um par conceitual (A, B), é calculado o valor do OP-RefD, prevendo se o conceito B é um pré-requisito de A ou não. Os pares em que A é um pré-requisito de B serão classificados como exemplos negativos.

Os experimentos foram conduzidos utilizando uma base de dados diferente do artigo original de Liang et al. (2015), a DBpedia ao invés da Wikipedia, sendo assim, para termos uma comparação válida na avaliação da modificação, os métodos EQUAL e TF-IDF foram também implementados para o cálculo do OP-RefD utilizando a estrutura entre conceitos da DBpedia.

Para o OP-RefD, foram combinadas as diferentes funções de peso e corte descritas no capítulo 4. A tabela 5 a seguir, mostra uma matriz com a combinação de todos os métodos experimentados com suas respectivas funções de peso e de corte utilizadas em cada um. O nome utilizado para cada um dos métodos servirá como referência nas tabelas comparativas com os resultados exibidos posteriormente neste trabalho.

Tabela 5: Relação dos métodos experimentados e suas respectivas funções de peso e corte

Método	Função de Peso	Função de Corte
Equal	EQUAL	N/A
TFIDF	TF-IDF	N/A
LevenPeso	Levenshtein	N/A

LevenNormPeso	Levenshtein Normalizado	N/A
SimilarPeso	Similaridade Word2Vec	N/A
TFIDF_Prop	TF-IDF Prop	N/A
EqualCorteSimilar	EQUAL	Similaridade Word2Vec
TFIDFCorteSimilar	TF-IDF	Similaridade Word2Vec
LevenCorteSimilar	Levenshtein	Similaridade Word2Vec
LevenNormCorteSimilar	Levenshtein Normalizado	Similaridade Word2Vec
SimilarCorteSimilar	Similaridade Word2Vec	Similaridade Word2Vec
TFIDF_PropCorteSimilar	TF-IDF Prop	Similaridade Word2Vec
EqualCorteLeven	EQUAL	Levenshtein
TFIDFCorteLeven	TF-IDF	Levenshtein
LevenCorteLeven	Levenshtein	Levenshtein
LevenNormCorteLeven	Levenshtein Normalizado	Levenshtein
SimilarCorteLeven	Similaridade Word2Vec	Levenshtein
TFIDF_PropCorteLeven	TF-IDF Prop	Levenshtein
EqualCorteLevenNorm	EQUAL	Levenshtein Normalizado
TFIDFCorteLevenNorm	TF-IDF	Levenshtein Normalizado
LevenCorteLevenNorm	Levenshtein	Levenshtein Normalizado
LevenNormCorteLevenNorm	Levenshtein Normalizado	Levenshtein Normalizado

SimilarCorteLevenNorm	Similaridade Word2Vec	Levenshtein Normalizado
TFIDF_PropCorteLevenNorm	TF-IDF Prop	Levenshtein Normalizado

Fonte: Própria.

Para melhor clareza, considere como um exemplo da tabela 5, o método "TFIDFCorteLeven". Ele utiliza como função de peso o TF-IDF e como função de corte a distância de Levenshtein. Nos experimentos realizados utilizou-se a distância de Levenshtein em duas versões diferenciadas, a primeira utilizando seu valor absoluto e a segunda utilizando seu valor normalizado.

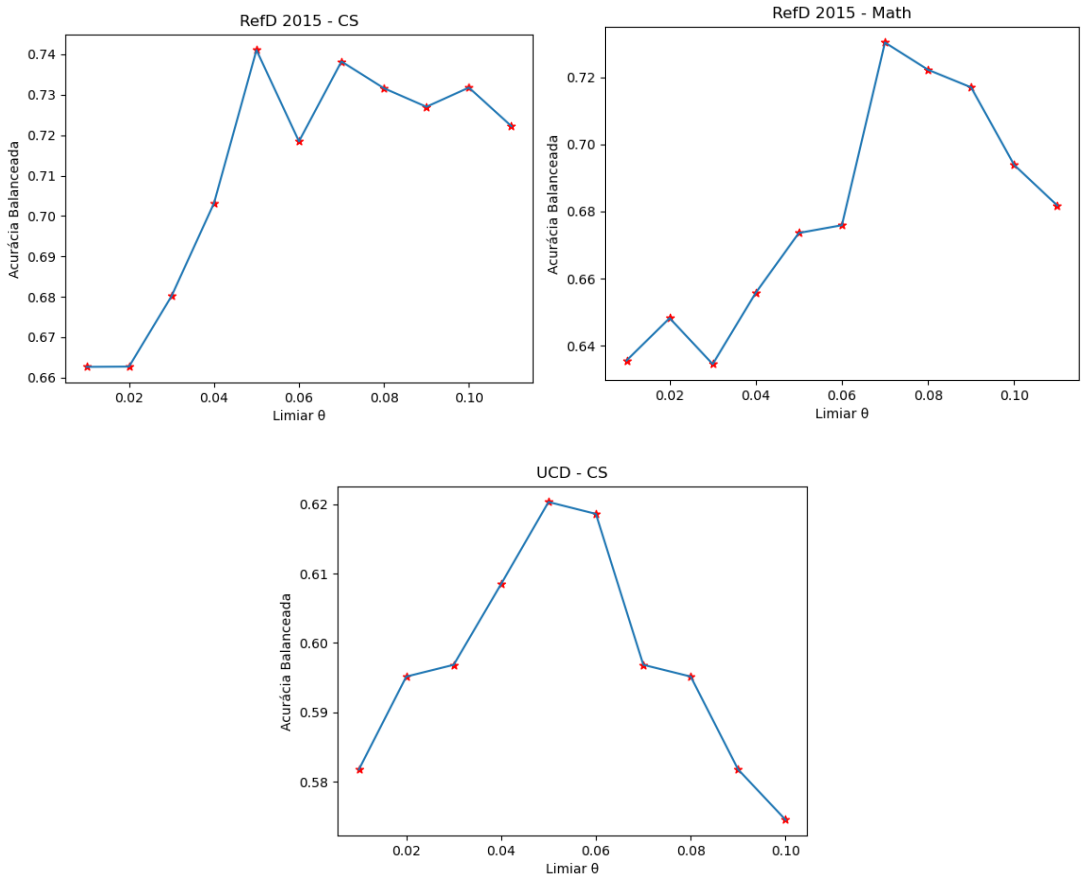
4.3.1. Parametrização do limiar de avaliação

Uma vez que o uso do RefD original para a previsão de pré-requisitos exige, conforme descrito ao final da seção 2.3.1, a definição de um limiar de avaliação θ, também foram executados experimentos para determinar o melhor valor de θ, investigando sua relação com o desempenho da previsão automática de pré-requisitos. A definição do θ foi feita com o cálculo do OP-RefD utilizando o método TF-IDF como função de peso, sem utilização de função de corte, uma vez que esta configuração serve como base[13] de melhor resultado encontrado nos experimentos do RefD original.

Os gráficos exibidos na figura 7 mostram a relação entre o limiar θ definido para avaliação e a acurácia média dos resultados para cada um dos *datasets* utilizados, subdivididos em domínio. Liang et al. (2015) citam que, empiricamente, um valor de limiar θ definido entre 0.02 e 0.1 produz um bom desempenho para a previsão de pré-requisitos

[13] Neste trabalho utilizaremos o termo "base" ou "linha de base" para nos referir ao que inglês costuma-se adotar a terminologia "baseline"

Figura 7: Relação entre o limiar θ do OP-RefD TF-IDF e a acurácia média nos três *datasets*



Fonte: Própria.

Desta forma, foi calculada a acurácia média para valores de θ entre 0.02 e 0.1 para cada um dos três *datasets*, e, conforme exibido nos gráficos da figura 8, para maximizar o desempenho da medida acurácia média, foi definida a utilização do valor θ em 0.05 para o *dataset* RefD2015 CS, 0.07 para o *dataset* RefD2015 Math, e 0.05 para o *dataset* UCD.

4.3.2. Parametrização das funções de corte

Os limiares para cada função de corte foram definidos experimentalmente de forma a maximizar a medida F1. O método utilizado como função de peso no cálculo do RefD também foi o TF-IDF, considerando os limiares para cada um dos *datasets* definidos conforme descrito na seção 4.3.1.

Lembrando que cada uma das medidas usadas como função de corte retorna uma pontuação de similaridade em uma determinada escala. A similaridade semântica Word2Vec e a distância *levenshtein* normalizada retornam um valor entre 0 e 1. Além disto, foi decidido manter os resultados do corte feito usando a distância *levenshtein* em valores absolutos do número de operações realizadas, pois os resultados preliminares se mostraram promissores.

Após testar vários valores de corte, foram definidos os limites para cada um dos limiares de forma a maximizar a medida da acurácia média de forma que os pares de resultados com a medida de similaridade acima do limite são considerados semelhantes. Assim, apenas os conceitos c_i mais semelhantes ao conceito A entraram no escopo do cálculo do OP-RefD.

Os limites escolhidos para cada medida são mostrados na tabela 6. Eles são os valores de corte utilizados para o cálculo do OP-RefD para cada *dataset* avaliado e função de corte desenvolvida. Em quaisquer dos métodos descritos na tabela 5 onde uma função de corte tenha sido utilizada, o corte considerado foi o valor conforme indicado na tabela 6.

Tabela 6: Valores de corte utilizados nas versões do OP-RefD para cada *dataset* e função de corte.

Dataset	**Função de Corte**	**Valor de Corte**
RefD2015 - *Math*	Similaridade Word2VEC	0.7
RefD2015 - *Math*	Levenshtein	5
RefD2015 - *Math*	Levenshtein Normalizada	0.2
RefD2015 - CS	Similaridade Word2VEC	0.9
RefD2015 - CS	Levenshtein	16

RefD2015 - CS	Levenshtein Normalizada	0.4
UCD	Similaridade Word2VEC	0.6
UCD	Levenshtein	12
UCD	Levenshtein Normalizada	0.9

Fonte: Própria.

A acurácia média foi utilizada para determinação dos limiares por se tratar de classes desbalanceadas (quantidade de exemplos positivos e negativos é diferente na classe real).

5. ANÁLISE DOS RESULTADOS

Este capítulo tem o objetivo de analisar os resultados encontrados no estudo experimental de avaliação dos diferentes métodos propostos como variações para funções de peso e de corte e aplicação no cálculo da distância de referência. Uma comparação em relação à performance do tempo de execução no cálculo do OP-RefD também será realizada.

5.1. Métricas de Avaliação dos Resultados

As medidas estatísticas que serão usadas para comparar as variações do OP-RefD serão a precisão, *recall* (revocação), medida F1, acurácia geral e acurácia média. Elas serão brevemente descritas aqui em seu emprego no contexto deste trabalho.

A precisão mede o número de vezes que a relação de pré-requisito existente foi corretamente prevista pelo método do OP-RefD (verdadeiros positivos), dividido pelo número total de vezes (verdadeiros positivos + falsos positivos) que a relação de pré-requisito foi prevista como existente no par conceitual. Para maximizar a precisão, é necessário que a relação de pré-requisito no par conceitual não seja incorretamente identificada como existente. A desvantagem desta métrica é que ela não leva em consideração os pares conceituais que deveriam ter sido identificados como relação de precedência não existente.

O *recall* (ou revocação) cobre essa desvantagem, pois mede, dentre todas as situações que o OP-RefD deveria ter classificado a precedência como existente (verdadeiro positivo + falso negativo), quantas estão corretas, pois considera as relações que o OP-RefD classificou como inexistente, quando existia.

A acurácia geral mede o quão efetivo o OP-RefD é do ponto de vista da classificação – relação de pré-requisito existe ou não existe, pois mede a taxa de acertos existente (verdadeiro positivo + verdadeiro negativo) com relação a todos os casos. O problema desta métrica é que ela não reflete a realidade quando os conjuntos de dados a serem avaliados não possuem a mesma quantidade de elementos por cada classe, no nosso caso, número de pares onde existe a relação de pré-requisitos e número de pares onde não existe a relação de pré-requisitos, que é justamente o caso dos *datasets* avaliados neste trabalho.

Além disto é importante citar a acurácia média (ou balanceada). Ela é mais adequada para distribuição desbalanceada das classes nos dados, como no trabalho atual, pois é calculada como a média da proporção corrigida de cada classe individualmente.

Finalmente, a medida F1 é a média harmônica entre a precisão e o recall e, desta forma, mede a eficiência do OP-RefD considerando os erros nas duas classes (falsos positivos e falsos negativos). Ou seja, para o F1 aumentar, é necessário que o acerto aconteça tanto na quantidade de relações corretamente identificadas como existentes quanto na quantidade corretamente identificadas como inexistentes.

Geralmente, para classes desbalanceadas, a escolha entre acurácia média e medida F1 depende de qual das duas classes (positiva ou negativa) supera a outra. Se o número de exemplos positivos é maior, o mais recomendado é utilizar a acurácia média, caso contrário, a medida F1.

Nos conjuntos de dados apresentados no capítulo 4, a quantidade de exemplos positivos fica aproximadamente entre 11% no *dataset* RefD 2015 – *Math* e 16% no dataset RefD 2015 – CS, indicando assim que nestes datasets o melhor é analisar os valores finais da medida F1. Enquanto que no *dataset* UCD a quantidade de exemplos positivos representa aproximadamente 60% dos casos, o que indica que o melhor é realizar a análise

da acurácia média.

É importante comentar também sobre as análises gráficas existentes nos problemas de previsão em modelos de classificação. Então será comentado aqui sobre as curvas ROC e as curvas de Precisão-Recall.

O gráfico das curvas ROC (*Receiver Operating Characteristic Curve*) ilustra o desempenho de um sistema classificador binário à medida que o seu limiar de discriminação varia. Estas curvas resumem bem o *trade-off* (relação de perda-ganho) entre a taxa positiva verdadeira e a taxa positiva falsa para um modelo preditivo usando diferentes limites de probabilidade.

Por outro lado, as curvas Precisão-Recall (PRC – *Precision Recall Curve*) resumem o *trade-off* (relação de perda-ganho) entre a taxa positiva verdadeira e o valor preditivo positivo para um modelo preditivo usando diferentes limites de probabilidade. Enquanto o *recall* identifica a taxa na qual os exemplos da classe positiva são previstas corretamente, a precisão indica a taxa na qual as previsões positivas estão corretas

Pelas próprias definições, as curvas ROC são apropriadas quando as observações entre cada classe são balanceadas, enquanto as curvas PRC são apropriadas para conjuntos de dados desequilibrados (SAITO; REHMSMEIER, 2015).

Por este motivo, as análises mais adiante neste capítulo serão feitas utilizando-se curvas de precisão-recall, uma vez que os três conjuntos de dados utilizados são desbalanceados.

5.2. Comparação entre os Resultados

O objetivo dos experimentos empíricos foi encontrar uma combinação válida de funções de peso e corte que alcançasse um melhor resultado em relação a (LIANG et al. 2015).

Além disto, pensando que o OP-RefD pode ser utilizado em ambientes online,

contexto que exige uma performance imediata e, considerando que o OP-RefD, enquanto medida, é um cálculo simples, com resultados rápidos, o objetivo também é encontrar uma combinação, não somente com bom desempenho nas medidas estatísticas em relação à linha de base (*baseline*) adotada, mas também com menor tempo de execução na definição da relação conceitual entre dois conceitos avaliados A e B.

As tabelas 7, 8 e 9 mostram os resultados encontrados para cada um dos três conjuntos de dados analisados. As medidas estatísticas na tabela são A – Acurácia, P – Precisão, R – Recall, F1 – Medida F1 e Tempo. Tempo corresponde ao tempo total de processamento do cálculo do OP-RefD para todos os pares conceituais avaliados de cada *dataset* e está expresso em segundos. Nestas tabelas, os valores em vermelho encontram-se em destaque pois são os métodos de peso adotados por Liang et al. (2015). Eles foram também implementados por este trabalho no contexto da DBpedia e seguem como base para as comparações.

Tabela 7: Relação dos métodos e métricas avaliadas para o *dataset* RefD2015 - Math

Método	A	P	R	F1	Tempo
Equal	0.6173	0.4510	0.8846	0.5974	5.01
TFIDF	0.6769	**0.5250**	0.9130	**0.6667**	**11.95**
LevenPeso	0.6267	0.4792	0.8846	0.6216	29.62
LevenNormPeso	0.5314	0.2500	0.7727	0.3778	29.63
SimilaridadePeso	0.6344	0.4746	0.9032	0.6222	71.71
TFIDF_PropPeso	0.6098	0.4423	0.8846	0.5897	11.32
EqualCorteSimilar	0.6267	0.4490	0.9565	0.6111	4.24
TFIDFCorteSimilar	**0.6825**	**0.5128**	**0.9524**	**0.6667**	**9.83**

LevenCorteSimilar	0.6438	0.4681	0.9565	0.6286	25.50
LevenNormCorteSimilar	0.6438	0.4681	0.9565	0.6286	25.36
SimilarCorteSimilar	0.6289	0.4516	0.9333	0.6087	64.34
TFIDF_PropCorteSimilar	0.6400	0.4583	0.9565	0.6197	9.27
EqualCorteLeven	0.6375	0.4706	0.9231	0.6234	3.99
TFIDFCorteLeven	**0.6923**	**0.5476**	**0.9583**	**0.6970**	**5.58**
LevenCorteLeven	0.6267	0.4792	0.8846	0.6216	27.43
LevenNormCorteLeven	0.6267	0.4792	0.8846	0.6216	27.48
SimilarCorteLeven	0.6277	0.4667	0.9032	0.6154	66.84
TFIDF_PropCorteLeven	0.6125	0.4423	0.9200	0.5974	8.34
EqualCorteLevenNorm	0.6173	0.4423	0.9200	0.5974	4.13
TFIDFCorteLevenNorm	**0.6719**	**0.5238**	**0.9565**	**0.6769**	**4.83**
LevenCorteLevenNorm	0.6267	0.4694	0.9200	0.6216	22.41
LevenNormCorteLevenNorm	0.6267	0.4694	0.9200	0.6216	22.33
SimilarCorteLevenNorm	0.6061	0.4194	0.8966	0.5714	55.79
TFIDF_PropCorteLevenNorm	0.5828	0.3478	0.8000	0.4848	7.47

Fonte: Própria.

Na tabela 7, os resultados destacados em negrito serão comentados por apresentarem uma melhora em relação às métricas estatísticas ou ao tempo de processamento do OP-RefD para o *dataset* RefD 2015 Math. Os métodos que apresentaram tempo de processamento menor, porém piora nas métricas estatísticas, não

foram considerados nas análises. Os métodos que se utilizaram da função de peso TF-IDF aparecem em todos os destaques.

Dos três métodos de corte propostos, o melhor desempenho foi do corte pelo valor absoluto da Distância de Levenshtein, "TFIDFCorteLeven", apresentando valores superiores ao método original sem corte, uma vez que sua precisão foi de 0.5476 (o original foi de 0.5250) e a medida F1 foi de 0.6970 (o original foi de 0.6667). Um ponto que merece destaque é o tempo de execução deste método, que foi reduzido em cerca de 53% se comparado ao método com função de peso TF-IDF, sem utilização de função de corte.

A tabela 8 exibe os resultados do OP-RefD para o *dataset* RefD2015 CS. Os valores destacados em negrito seguem o mesmo padrão de marcação mencionada anteriormente – para a tabela 7.

Tabela 8: Relação dos métodos e métricas avaliadas para o *dataset* RefD2015 - CS

Método	A	P	R	F1	Tempo
Equal	0.7214	0.6386	0.8548	0.7310	6.67
TFIDF	0.7311	**0.6575**	0.8727	**0.7500**	**13.35**
LevenPeso	0.7023	0.6203	0.8448	0.7153	35.20
LevenNormPeso	0.5864	0.4505	0.7353	0.5587	35.28
SimilaridadePeso	0.7114	0.6477	0.8261	0.7261	78.96
TFIDF_PropPeso	0.6986	0.6250	0.8333	0.7143	14.45
EqualCorteSimilar	0.7214	0.6429	0.8571	0.7347	5.34
TFIDFCorteSimilar	0.7227	0.6622	0.8596	**0.7481**	**12.04**

LevenCorteSimilar	0.7023	0.6250	0.8475	0.7194	33.42
LevenNormCorteSimilar	0.7023	0.6250	0.8475	0.7194	33.53
SimilarCorteSimilar	0.7020	0.6304	0.8406	0.7205	74.92
TFIDF_PropCorteSimilar	0.6959	0.6180	0.8333	0.7097	12.32
EqualCorteLeven	0.7388	0.6410	0.8772	0.7407	5.15
TFIDFCorteLeven	0.7727	**0.7077**	0.8846	**0.7863**	**5.65**
LevenCorteLeven	0.7188	0.6234	0.8727	0.7273	24.79
LevenNormCorteLeven	0.7188	0.6234	0.8727	0.7273	24.81
SimilarCorteLeven	0.7111	0.6250	0.8475	0.7194	45.72
TFIDF_PropCorteLeven	0.7222	0.6265	0.8525	0.7222	9.22
EqualCorteLevenNorm	0.7092	0.6235	0.8548	0.7211	5.11
TFIDFCorteLevenNorm	0.7373	**0.6528**	0.8868	**0.7520**	**6.02**
LevenCorteLevenNorm	0.7045	0.6125	0.8596	0.7153	31.36
LevenNormCorteLevenNorm	0.7045	0.6125	0.8596	0.7153	31.36
SimilarCorteLevenNorm	0.7042	0.6163	0.8548	0.7162	71.00
TFIDF_PropCorteLevenNorm	0.6770	0.5938	0.8143	0.6867	8.42

Fonte: Própria.

Pode-se observar na tabela 8 que os métodos com melhor desempenho também foram aqueles que combinaram a função de peso TF-IDF com as diferentes funções de corte. A função de corte com melhor desempenho foi a do valor absoluto da Distância de Levenshtein, que considerou valores abaixo de 16 operações, apresentando valores

superiores ao método original sem corte, uma vez que sua precisão foi de 0.7077 (o original foi de 0.6575). Em relação à medida F1, o método desenvolvido aqui teve um valor de 0.7863, frente ao F1 do método original sem corte, que foi de 0.7500.

Mais uma vez, deve-se destacar que o tempo de execução deste método foi reduzido em aproximadamente 58% se comparado com o método com função de peso TF-IDF, sem utilização de função de corte.

O método com função de peso TF-IDF e função de corte utilizando similaridade Word2Vec teve um tempo de execução comparável à linha de base, porém teve seu desempenho piorado em relação à medida F1.

Observa-se na tabela 9 os resultados do OP-RefD para o *dataset* UCD. Nesta tabela foi adicionado o resultado da BA – Acurácia Balanceada, ou média. Os valores marcados em negrito são os destaques encontrados neste conjunto de dados.

Tabela 9: Relação dos métodos e métricas avaliadas para o *dataset* UCD

Método	A	P	R	F1	BA	Tempo
Equal	0.6940	0.7997	0.7843	0.7919	0.6088	26.24
TFIDF	0.7050	**0.8200**	0.7836	**0.8014**	**0.6203**	51.00
LevenPeso	0.6919	0.8024	0.7783	0.7902	0.6087	110.35
LevenNormPeso	0.5824	0.7128	0.6398	0.6743	0.5513	110.37
SimilaridadePeso	0.6640	0.7800	0.7516	0.7655	0.5895	250.49
TFIDF_PropPeso	0.6951	0.8023	0.7808	0.7914	0.6156	48.30
EqualCorteSimilar	0.6871	0.7925	0.7850	0.7887	0.5939	19.59
TFIDFCorteSimilar	0.7028	0.8126	0.7910	**0.8017**	0.6076	**45.38**

LevenCorteSimilar	0.7029	0.8007	0.7980	0.7993	0.6140	80.45
LevenNormCorteSimilar	0.7029	0.8007	0.7980	0.7993	0.6140	80.50
SimilarCorteSimilar	0.6716	0.7808	0.7661	0.7734	0.5903	201.69
TFIDF_PropCorteSimilar	0.6867	0.7905	0.7880	0.7892	0.5898	36.81
EqualCorteLeven	0.6979	0.7984	0.7946	0.7965	0.6056	18.69
TFIDFCorteLeven	0.7090	**0.8201**	0.7949	**0.8073**	**0.6109**	**25.55**
LevenCorteLeven	0.6949	0.7997	0.7878	0.7937	0.6055	92.46
LevenNormCorteLeven	0.6949	0.7997	0.7878	0.7937	0.6055	92.46
SimilarCorteLeven	0.6731	0.7831	0.7636	0.7732	0.5961	200.47
TFIDF_PropCorteLeven	0.6968	0.7991	0.7916	0.7953	0.6062	32.51
EqualCorteLevenNorm	0.6940	0.7997	0.7843	0.7919	0.6088	19.50
TFIDFCorteLevenNorm	0.6975	**0.8225**	0.7721	**0.7965**	**0.6123**	**24.04**
LevenCorteLevenNorm	0.6919	0.8024	0.7783	0.7902	0.6087	102.55
LevenNormCorteLevenNorm	0.6919	0.8024	0.7783	0.7902	0.6087	102.73
SimilarCorteLevenNorm	0.6644	0.7813	0.7516	0.7661	0.5892	244.02
TFIDF_PropCorteLevenNorm	0.6423	0.7714	0.7254	0.7477	0.5712	19.35

Fonte: Própria.

Analisando os resultados da tabela 9, mais uma vez os métodos com melhor desempenho são os que combinam a função de peso TF-IDF com as diferentes funções de corte apresentadas. Dentre as funções de corte, seguindo a tendência dos dois *datasets* anteriores, a que se destacou com melhor desempenho foi o corte da distância de

Levenshtein (corte com distância abaixo de 12), apresentando valores comparáveis ao método original sem corte, uma vez que sua precisão foi de 0.8201 enquanto o original foi de 0.8200. Na medida F1, o método desenvolvido aqui também não apresentou melhora considerável, pois teve um valor de 0.8073, frente ao F1 do método original sem corte, que foi de 0.8014. O mesmo ocorre na acurácia média, que ficou em 0.6109, abaixo da acurácia média do método original, no valor de 0.6203.

Para este conjunto de dados, o ganho nos métodos desenvolvidos também foi no tempo de execução, que se reduziu em aproximadamente 50%, comparando ao método com função de peso TF-IDF, sem utilização de função de corte.

O método com função de peso TF-IDF e função de corte utilizando similaridade Word2Vec teve um tempo de execução levemente melhorado em relação à linha de base, porém teve seu desempenho comparável ao original, em relação à medida F1.

5.3. Discussão geral dos resultados

Os resultados anteriores evidenciam, conforme literatura, que é possível identificar automaticamente os relacionamentos de pré-requisito entre conceitos na maioria dos casos. Pode-se acrescentar ainda que os resultados descritos mostram uma leve melhora nas medidas estatísticas utilizando o OP-RefD em relação ao original, indicando que ainda há espaço para explorar outras relações existentes, tanto usando a DBpedia como fonte de dados, quanto usando outros diferentes grafos de conhecimento.

No entanto, vale a pena investigar os casos em que a estratégia de cálculo do OP-RefD falha em identificar o relacionamento de pré-requisito entre conceitos. Este cenário pode ser explicado pelo fato de a estratégia automática de identificação de pré-requisitos depender muito dos conceitos existentes na fonte de onde foram extraídos, neste caso, a DBpedia. Isto também explica os diferentes valores de medidas estatísticas encontrados em relação ao trabalho original de Liang et al. (2015), utilizando os mesmos métodos de

peso originalmente propostos.

Um outro cenário de importante análise são os casos falsos positivos e negativos, aqueles em que o cálculo do OP-RefD falhou em dizer que havia ou não havia relação de pré-requisito no par conceitual. Este resultado era esperado devido ao uso de uma estratégia automática para a identificação de pré-requisitos que depende muito dos conceitos propostos e das fontes de onde foram extraídos (ou seja, DBpedia).

Esta análise pode ser melhor percebida ao verificarmos que, de maneira geral, em todos os métodos aplicados, a precisão do conjunto de dados do domínio de Matemática (RefD2015 *Math*) é mais baixa em relação aos outros dois conjuntos de dados no domínio de Ciência da Computação (RefD2015 CS e UCD). Isto pode ser explicado pois a cobertura de conceitos existentes na DBpedia para diferentes domínios pode variar muito. Uma vez que a DBpedia extrai o conteúdo estruturado da Wikipedia, deve-se considerar que nela alguns domínios são mais populares que outros – editados com mais frequência – contribuindo para uma melhor qualidade dos dados e uma estrutura de referências mais completa.

Em relação às estratégias de corte adotadas, a motivação foi que as informações adicionais encontradas no grafo de conhecimento fossem exploradas de forma a incorporar no cálculo do OP-RefD apenas os conceitos altamente relacionados e descartar os que atrapalhavam mais que ajudavam.

Os resultados mostram que o método do OP-RefD usando a função de peso TF-IDF e a função de corte pela distância de Levenshtein no geral, para os três *datasets* avaliados, produziram os melhores resultados de desempenho em relação às métricas estatísticas.

Para melhor visualizar esta conclusão e comparar a precisão e o *recall*, foram plotadas as curvas Precisão-Recall do método RefD TF-IDF original, e os métodos do

OP-RefD com função de peso TF-IDF e diferentes funções de corte. Uma medida interessante que pode-se calcular tanto a partir de curvas ROC, quanto a partir de uma curva precisão-*recall* chama-se área sobre a curva (AUC – *Area Under Curve*). Ela pode ser definida com um resumo estatístico que indica que quanto maior a AUC, maior o desempenho retornado por um classificador, ou seja, ele pode ser utilizado para comparar a performance de classificadores (MOTTA, 2016).

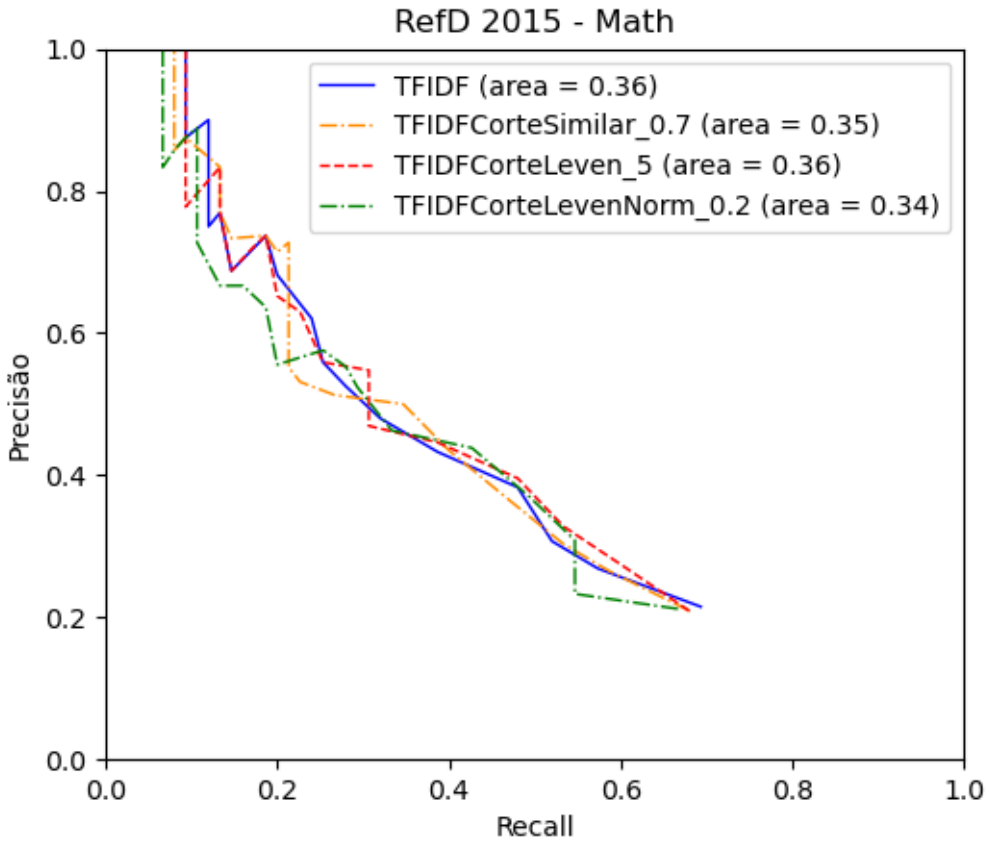
Em termos de verificação de valores para a comparação de performance, a linha de base do valor AUC da curva Precisão-Recall não é fixa como no caso do AUC para curvas ROC. Ela é determinada pela razão de positivos (P) e negativos (N), ou seja, o valor base da AUC de um PCR = P / (P + N) (SAITO; REHMSMEIER, 2015). Se o valor do AUC retornado é igual a 1 então pode-se dizer que o classificador é perfeito e caso o AUC seja igual à linha de base, o classificador é considerado randômico. Na prática os classificadores devem estar entre 1 e o valor de linha de base.

Assim, para correta interpretação dos gráficos exibidos nas figuras 8, 9 e 10, é importante sabermos os valores base para cada um dos *datasets*. Para o conjunto de dados do RefD15 Math o valor base do AUC ficou em 0.11. Para o RefD15 CS em 0.16 e para o UCD em 0.60.

Os gráficos das figuras 8, 9 e 10 foram construídos a partir dos valores de precisão e recall obtidos ao realizar o cálculo do OP-RefD com diferentes valores do limiar de avaliação θ, conforme explicado na seção 4.3.1. Os experimentos foram executados com valor θ entre 0.01 e 0.29 e para cada *dataset* há um ponto onde se atinge a precisão máxima, ponto este em que o valor de recall começa a decrescer. Para cada *dataset* o gráfico foi plotado considerando até o valor de precisão máxima. Valores acima deste ponto não foram considerados, pois após este valor máximo de θ, a previsão do modelo já não consegue acertar como esperado.

A figura 8 exibe o gráfico para o *dataset* RefD 2015 – Math. Pode-se observar a área abaixo da curva para cada um dos métodos plotados, exibindo o valor da precisão média.

Figura 8: Comparação das curvas Precisão-Recall do RefD original e o OP-RefD com diferentes cortes para o *dataset* RefD 2015 - Math



Fonte: Própria.

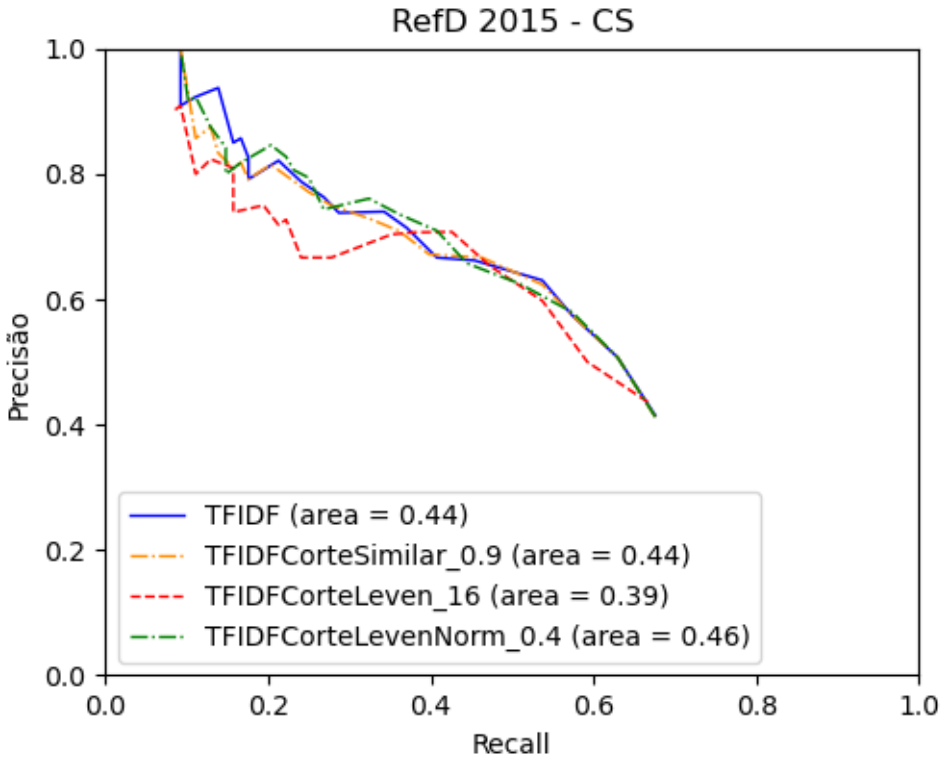
Com o método utilizando função de peso TF-IDF e função de corte Levenshtein desempenhando com a precisão média de 0.36, igualando o método original TF-IDF. Os demais métodos do OP-RefD não tiveram resultado expressivo acima do original.

Considerando que o valor base do AUC para este *dataset* é 0.11, todos os valores encontrados mostram uma boa resposta na previsão, para todos os métodos plotados. Os valores de limiar θ considerados para o Ref15 Math variou de 0.01 até no máximo 0.30, com passos de 0.01.

A figura 9 exibe o gráfico para o *dataset* RefD 2015 – CS. Observa-se o valor da

precisão média exibido no gráfico como o valor da área abaixo da curva para cada um dos métodos plotados.

Figura 9: Comparação das curvas Precisão-Recall do RefD original e o OP-RefD com diferentes cortes para o *dataset* RefD 2015 - CS



Fonte: Própria.

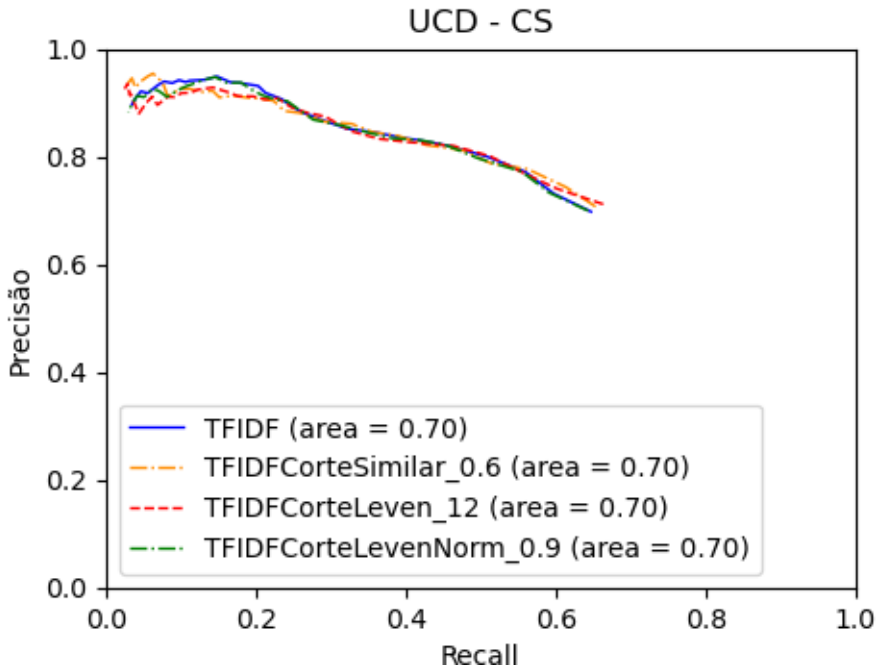
O método utilizando função de peso TF-IDF e função de corte Levenshtein Normalizado teve uma precisão média de 0.46, enquanto o método original TF-IDF ficou em 0.44. Os demais métodos do OP-RefD não tiveram grandes alterações na precisão média em relação ao método original.

No entanto, considerando que o valor base do AUC para este *dataset* é 0.16, todos os valores encontrados estão acima deste valor, o que comprova uma boa resposta na previsão dos pré-requisitos entre os conceitos, para todos os métodos plotados. Para o conjunto de dados Ref15 CS, os valores de limiar θ considerados para o Ref15 CS variou de 0.01 até no máximo 0.30, com passos de 0.01.

Na figura 10, pode-se visualizar o gráfico para o *dataset* UCD. O valor da precisão

média, representado pelo valor AUC da área abaixo da curva, é plotado para cada um dos métodos.

Figura 10: Comparação das curvas Precisão-Recall do RefD original e o OP-RefD com diferentes cortes para o *dataset* UCD



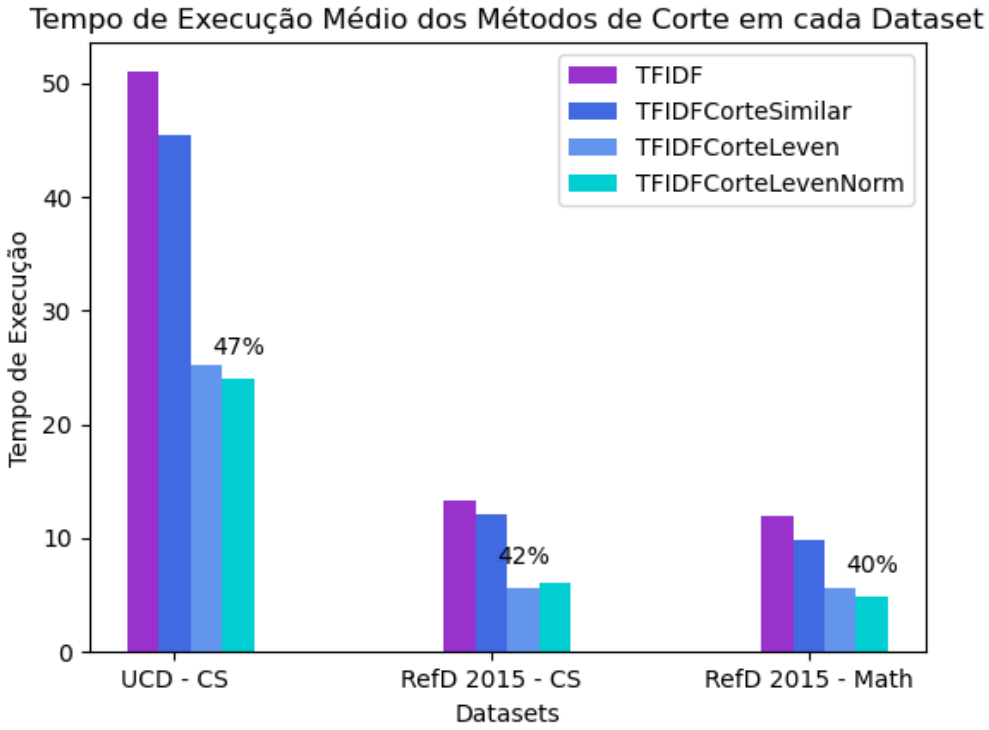
Fonte: Própria.

Nenhum dos métodos que se utilizaram da função de peso TF-IDF e das funções de corte variadas desempenhou melhor na precisão média que o método original TF-IDF, todos ficando com valores muito próximos, para cada um dos métodos.

Porém, confrontando o valor base do AUC para este *dataset* de 0.60, todos os valores encontrados de AUC caracterizam uma boa resposta na previsão, para todos os métodos plotados. Para o conjunto de dados UCD CS, os valores de limiar θ considerados variou de 0.01 até no máximo 0.30, com passos de 0.01.

A figura 11 exibe o gráfico comparativo do tempo de processamento médio de cada um dos *datasets* avaliados. Para cada *dataset* é exibido o percentual comporativo entre o método com melhor desempenho em relação ao tempo de execução do método original TF-IDF.

Figura 11: Comparação dos tempos de execução do RefD original e o OP-RefD com diferentes cortes para os *datasets*



Fonte: Própria.

Pode-se observar que o melhor desempenho, com menor termpo de processamento médio, foi o OP-RefD com método TF-IDF e corte Leveishtein normalizado para os *datasets* UCD e RefD 2015 – Math. Por outro lado, o melhor desempenho para o *dataset* RefD 2015 – CS foi o OP-RefD com método TF-IDF e corte Leveishtein (valor absoluto).

Conforme mencionado anteriormente, o ganho de desempenho no tempo de execução dos métodos de identificação de pré-requisitos variou entre 50% a 58%.

6. CONCLUSÃO

Neste capítulo serão apresentadas as conclusões finais desta dissertação, descrevendo as principais contribuições, limitações e trabalhos futuros.

6.1. Considerações finais

Esta dissertação mostrou o desenvolvimento de um método adaptado para identificação automática de pré-requisitos entre conceitos. Além disto, foi feito um estudo experimental e uma comparação entre uma medida para a identificação dos pré-requisitos entre conceitos da literatura – RefD (distância de referência) – e sua versão modificada – OP-RefD (distância de referência – performance otimizada) –, apresentada neste trabalho. O processo de identificação da relação de pré-requisitos explorou tanto relações sintáticas de similaridade entre strings, quanto relações semânticas entre conceitos, que podem ser encontradas em um grafo de conhecimento. O estudo comparativo considerou como linha de base (*baseline*) o cálculo da medida RefD original, porém aplicada no contexto da DBpedia.

Foram discutidas as diferentes estratégias de peso utilizadas ao calcular a relação entre os conceitos do par avaliado. Além disto foram utilizadas diferentes funções de corte e assim, foi possível reduzir o espaço conceitual do grafo de conhecimento a um conjunto menor de conceitos relacionados aos conceitos do par avaliado, buscando uma melhora na identificação automática.

O processo de avaliação dos métodos implementados foi realizado em diferentes etapas. Na etapa de configuração dos parâmetros, foi considerado o método da linha de base para realização do ajuste do limiar de avaliação do cálculo do RefD Original ideal

no contexto da DBpedia. Uma segunda configuração envolveu a definição de parâmetros das diferentes funções de corte, onde cada algoritmo foi avaliado variando as configurações do parâmetro de corte para cada função proposta.

A terceira etapa avaliou os métodos modificados utilizando as métricas estatísticas mais adequadas ao contexto de cada um dos três *datasets* avaliados para verificar se houve melhora no desempenho em algum dos métodos.

6.2. Contribuições

A principal contribuição dessa pesquisa foi o desenvolvimento de um método adaptado (OP-RefD) para identificação automática de pré-requisitos entre conceitos. Além do planejamento, a execução e a análise dos resultados de um estudo experimental para comparar os métodos propostos com o existente na literatura.

Adicionalmente foi possível demonstrar que a identificação de relações de pré-requisito entre pares de conceitos pode ser aplicada utilizando a estrutura de um grafo de conhecimento aberto, tal qual a DBpedia, sem perda de eficácia (medidas estatísticas) e com eficiência (tempo de processamento) maior em relação à base de comparação.

Além disto, as principais contribuições técnicas desta Dissertação são:

- A definição e implementação de métodos para identificação de pré-requisitos entre pares conceituais, com diferentes funções de peso e funções de corte.
- Criação de ambiente em nuvem, disponível publicamente[14], com instalação do Virtuoso e versão das triplas RDF da DBpedia 2016-10.

Os algoritmos foram implementados em Python e se encontram publicamente

[14] Disponível no Painel EC2 da AWS, buscando AMI por "DBPEDIA EN CORE sem NIF com Python3"

disponíveis no repositório do GitHub[15]. No Apêndice A, se encontram as consultas SPARQL desenvolvidas e submetidas na instalação da DBpedia.

6.3. Resposta para a questão de pesquisa

Após a análise dos resultados do estudo experimental apresentado no capítulo 5, é possível responder à questão de pesquisa proposta neste trabalho:

RQ1: O método adaptado proposto encontra melhores resultados em menor tempo de execução quando comparado ao método original ao ser aplicado em uma estrutura de dados (grafo de conhecimento) mais amplo em termos de existencia de mais relações estruturais entre conceitos?

Para os três conjuntos de dados analisados, o método OP-RefD obteve resultados estatísticos similares em relação à linha de base e tempo de processamento menor comparado à base, variando com um ganho entre 40% e para o *dataset* que desempenhou pior (RefD2015 Math) e 47% para o *dataset* que desempenhou melhor (UCD).

Assim, conclui-se que o método proposto tem sua eficácia comparável ao método original em relação às métricas estatísticas e desempenho melhorado em termos de processamento, uma vez que tem tempos de execução inferiores para os três conjuntos de dados avaliados.

6.4. Limitações

Uma importante limitação da abordagem deste trabalho é que os conceitos dos pares de pré-requisito avaliados devem estar no espaço de conceitos do grafo de conhecimento. No caso da implementação neste trabalho, os conceitos devem estar na DBpedia. Outro ponto relevante a ser mencionado é o fato de os conjuntos de dados

[15] https://github.com/rubiasa/OP-RefD

testados serem limitados a apenas dois domínios diferentes (computação e matemática).

Outra limitação é que não há garantias de que os pré-requisitos mais relevantes sejam recuperados no processo de corte, por se tratar de um processo automático, com limites definidos experimentalmente, sempre há o risco de falsos positivos, conforme observado nos resultados obtidos.

6.5. Trabalhos futuros

Os possíveis trabalhos futuros sugerem que a estrutura da DBpedia seja explorada mais em sua estrutura, expandindo o espaço de busca dos conceitos em comum para além dos conceitos vizinhos no grafo e sim considerando conceitos com caminho de tamanho acima de 1 e aplicando outros métodos de corte por similaridade semântica deixando apenas os conceitos mais fortemente relacionados ao conceito em avaliação

Outra possibilidade neste sentido é explorar as relações de categorias dos conceitos, uma vez que todo artigo na Wikipedia, e consequentemente, todo conceito na DBpedia, pertence a uma ou mais categorias, um caminho seria verificar como a relação destas categorias dos conceitos relacionados a um dos conceitos avaliados influenciaria nas funções de peso e nas funções de corte.

Há também a possibilidade de análise de diferentes tipos de relacionamentos e suas semânticas no grafo da DBpedia.

REFERÊNCIAS BIBLIOGRÁFICAS

ANTONIOU, Grigoris; HARMELEN, Frank v. A Semantic Web Primer. 2.ed. **Cooperative Information Systems**. England: MIT Press, 2008.

AUSUBEL, David P. The psychology of meaningful verbal learning. United States: **New York Grune & Stratton**, 1963.

BERGAN, John; JESKA, Patrick. An examination of prerequisite relations, positive transfer among learning tasks, and variations in instruction for a seriation hierarchy. **Contemporary Educational Psychology**. v. *5,* n. 3, p. 203-215, 1980.

BERNERS-LEE, Tim; HENDLER, James; LASSILA, Ora. The semantic web. **Scientific American**. v.284, n.5, p.34-43, 2001.

BICALHO, Maria Gabriela Parenti; MORAIS, Rossana Cristina Ribeiro. Ciberespaço e território: construção de uma discussão interdisciplinar. **PerCursos,** v. 17, n. 34, p. 05-23, 2016.

BOLLACKER, Kurt; EVANS, C; PARITOSH, P; STURGE, T.; TAYLOR, J. Freebase: a collaboratively created graph database for structuring human knowledge. In: **Proceedings of the 2008 ACM SIGMOD international conference on Management of data**. p. 1247-1250, 2008.

CHANGUEL, Sahar, LABROCHE, Nicolas, BOUCHON-MEUNIER, Bernadette. Resources Sequencing Using Automatic Prerequisite –Outcome Annotation. **ACM Transaction Intelligent Systems and Technology (TIST)**. v. 6*,* n. 1, p.6, 2015.

SCHMACHTENBERG, Max, BIZER, Christian, JENTZSCH, Anja, CYGANIAK, Richard. Linking open data cloud diagram. **The Linking Open Data cloud diagram**, p. 08-30, 2014.

DE MEDIO, Carlo; GASPARETTI, Fabio; LIMONGELLI, Carla; SCIARRONE, Filippo et al. Automatic extraction and sequencing of Wikipedia Pages for smart course building. In: **21st International Conference Information Visualization (IV).** IEEE, Londres, Inglaterra, 2017.

DIAS, Tatiane Domingos. Web Semântica: Conceitos Básicos e Tecnologias Associadas. **Web Semântica: Fundamentos e Tecnologias.** Universidade do Estado do Rio de Janeiro, 2001

FILLMORE, Charles J. Frame semantics. **Cognitive linguistics: Basic readings**. n. 34, pp.373-400, 2006.

FONSECA, João J. S. Metodologia da pesquisa científica. Fortaleza: **UEC**, 2002.

GAO, Li; DAI, Kun; GAO, Liping; JIN, Tao. Expert knowledge recommendation systems based on conceptual similarity and space mapping. **Expert Systems with Applications**. v. 136, pp. 242-251, 2019.

GASPARETTI, Fabio; LIMONGELLI, Carla; SCIARRONE, Filippo. Exploiting Wikipedia for Discovering Prerequisite Relationships Among Learning Objects. **International Conference Information Technology Based Higher Education and Training (ITHET)**. Lisboa, Portugal, 2015.

GIL, Antônio C. Como elaborar projetos de pesquisa. 4.ed. São Paulo: **Atlas**, 2007.

GOLDIE, John. G. S.. Connectivism: A knowledge learning theory for the digital age? **Medical teacher**. v. 38, n. 10, pp. 1064-1069, 2016.

HOFFART, Johannes; SUCHANEK, Fabian M.; BERBERICH, Klaus; WEIKUM, Gerhard. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. **Artificial Intelligence**, v. 194, p. 28-61, 2013.

HONNIBAL, Matthew; MONTANI, Ines. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. **To appear**. v. 7, n. 1, 2017.

HOUAISS, A. Dicionário Eletrônico Houaiss de Língua Portuguesa. Instituto Antônio Houaiss de Lexicografia e Banco de Dados da Língua Portuguesa S/C Ltda. Rio de Janeiro: **Objetiva**, 2009.

IEEE. 1484.12.1-2002 – IEEE Standard for Learning Object Metadata. **Institute of Electrical and Electronic Engineers – IEEE**. Published Date: 2002-09-06. Reaffirmed:2009-05-13. pp.1-40, 2002.

KIVUNJA, Charles. Exploring the pedagogical meaning and implications of the 4Cs "super skills" for the 21st century through Bruner's 5E lenses of knowledge construction to improve pedagogies of the new learning paradigm. **Creative Education**. v. 6, n. 02, 224, 2015.

LALA, Raja; GEEST, Marcell V.; RUSETI, Stefan; JEURING, Johan; DASCALU, Mihai; DORTMONT, Jordy V.; GUTU-ROBU, Gabriel; HULSBERGEN, Michiel. Enhancing Free-text Interactions in a Communication Skills Learning Environment. In: LUND, K., NICCOLAI, G. P., LAVOUÉ, E., GWEON, C. H., & BAKER, M. (Eds.). A Wide Lens: Combining Embodied, Enactive, Extended, and

Embedded Learning in Collaborative Settings. **13th International Conference on Computer Supported Collaborative Learning (CSCL)**. v. 2. pp. 877-878. Lyon, France, 2019.

LASLIE, M. The People's Encyclopedia. **Science**. v. 301. p. 1299, 2003.

LAURENCE, Stephen; MARGOLIS, Eric. Concepts and cognitive science. **Concepts: core readings**. v. 3, p. 81, 1999.

LEHMANN, Jens; ISELE, Robert; JAKOB, Max; JENTZSCH, Anja; KONTOKOSTAS, Dimitris; MENDES, Pablo; HELLMANN, Sebastian; MORSEY, Mohamed; VAN KLEEF, Patrick; AUER, Soren; BIZER, Christian. DBpedia–A large-scale, multilingual knowledge base extracted from Wikipedia. **Semantic Web**. v. 6, n. 2, p. 167–195. 2015.

LEVENSHTEIN, Vladimir I. Binary codes capable of correcting deletions, insertions, and reversals. **Soviet Physics Doklady**. v. 10. n. 8. pp. 707-710. 1966.

LÉVY, Pierre. Cyberculture. United States: **Univ of Minnesota Press**, 2001.

LIANG, Chen; YE, Jianbo; WU, Zhaohui; PURSEL, Bart; GILES, C. Lee. Recovering Concept Prerequisite Relations from University Course Dependencies. **Seventh Symposium on Educational Advances in Artificial Intelligence (EAAI)**. California, Estados Unidos, 2017.

LIANG, Chen; WU, Zhaohui; HUANG, Wenyi; GILES, Clyde L. Measuring prerequisite relations among concepts. In: **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**. Lisboa, Portugal, 2015.

LIMONGELLI, Carla; GASPARETTI, Fabio; SCIARRONE, Filippo. Wiki

course builder: a system for retrieving and sequencing didactic materials from wikipedia. In: **International Conference on Information Technology Based Higher Education and Training (ITHET).** IEEE, Lisboa, Portugal, 2015.

MANRIQUE, Rubén; SOSAY, Juan; MARINO, Olga; NUNES, Bernardo P; CARDOZO, Nicolas. Investigating learning resources precedence relations via concept prerequisite learning. **Web Intelligence (WI), IEEE WIC ACM International Conference on Santiago**. Santiago, Chile, 2018.

MANRIQUE, Rubén; PEREIRA, Bernardo; MARINO, Olga; CARDOZO, Nicolas; WOLFGAND, Sean. Towards the identification of concept prerequisites via Knowledge Graphs. In: **International Conference on Advanced Learning Technologies and Technology-enhanced Learning da The IEEE Computer Society e The IEEE Technical Committee on Learning Technology (ICALT)**. Alagoas, Brasil, 2019.

MEDELYAN, Olena; MILNE, David; LEGG, Catherine; WITTEN, Ian. H. Mining meaning from Wikipedia. **International Journal of Human-Computer Studies**. v. 67, pp. 716–754, 2009.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient estimation of word representations in vector space. ***arXiv preprint***. *arXiv*:1301.3781, 2013.

MILLER, George A, 1995. WordNet: a lexical database for English. **Communications of the ACM**. v. 38, n. 11, pp. 39-41, 1995

MINTZ, Mike; BILLS, Steven; SNOW, Rion; JURAFSKY, Daniel. Distant supervision for relation extraction without labeled data. In: **Proceedings of the Joint**

Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP. Suntec, Singapore, 2009.

MOTTA, Porthos R. A. Estudo exploratório do uso de classificadores para a predição de desempenho e abandono em universidade. **Dissertação (Mestrado em Ciência da Computação) - INF Instituto de Informática - Regional Goiânia**. Orientadora: Profª Drª Ana Paula Laboissière Ambrósio. Goiânia, 2016.

OHLAND, Matthew W.; YUHASZ, Amy G.; SILL, Benjamin L. Identifying and removing a calculus prerequisite as a bottleneck in Clemson's General Engineering Curriculum. **Journal of Engineering Education**. v. 93, n. 3, pp. 253-257, 2004.

PAN, Liangming; LI, Chengjiang.; LI, Juanzi.; TANG, Jie. Prerequisite relation learning for concepts in moocs. In: **55th Annual Meeting of the Association for Computational Linguistics**. Vancouver, Canada, 2017.

PAWAR, Atish S. Semantic similarity between words and sentences using lexical database and word embeddings. 2018. Dissertation (Master of Science in Computer Science) - **Lakehead University**, Thunder Bay, Canada, 2018.

PÉREZ, Jorge; ARENAS, Marcelo; GUTIERREZ, Claudio. Semantics and Complexity of SPARQL. **ACM Transaction on Database Systems**. v. 34, n. 3, 2009.

PRADO, Rafael de Lima. Armazenamento Otimizado de Dados RDF em um SGBD Relacional. 82 f. **Dissertação (Mestrado em Informática) - Universidade Federal do Paraná.** Orientadora: Profª Drª Carmem Satie Hara. Curitiba, 2017.

PRUD'HOMMEAUX, Eric; SEABORNE, Andy. SPARQL Query Language for RDF. Disponível em: <http://www.w3.org/TR/rdf-sparql-query/>. Acesso: 15.abr.2020,

2008.

ROY, Sudeshna; MADHYASTHA, Meghana; LAWRENCE, Sheril; RAJAN, Vaibhav. Inferring Concept Prerequisite Relations from Online Educational Resources. **The Thirty-First AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-19)**. Hawaii, Estados Unidos, 2019.

ROY, Devshri; SARKAR, Sudeshna; GHOSE, Sujoy. Automatic extraction of pedagogic metadata from learning content. **International Journal of Artificial Intelligence in Education**. v. 2, n. 18, pp. 97-118, 2008.

SAITO, Takaya; REHMSMEIER, Marc. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **PloS one**, v. 10, n. 3, 2015

SAYYADIHARIKANDEH, Mohsen; AMBIT, José L; LERMAN, Kristina; GORDON, Jonathan. Finding Prerequisite Relations using the Wikipedia Clickstream. **International World Wide Web Conference**. California, Estados Unidos, 2019.

SCHMITT, Xavier; KLUBER, Sylvain; ROBERT, Jérémy; PAPADAKIS, Mike; LETRAON, Yves. A Replicable Comparison Study of NER Software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate. In: **Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. Granada, Espanha, 2019.

TALUKDAR, Partha P.; COHEN, William W. Crowdsourced Comprehension: Predicting Prerequisite Structure in Wikipedia. **The 7th Workshop on the Innovative Use of NLP for Building Educational Applications**. Montreal, Canada, 2012.

TRIVIÑOS, AUGUSTO N. S. Introdução à pesquisa em ciências sociais: a

pesquisa qualitativa em educação. São Paulo: **Atlas**, 1987.

VAKARE, Tanmay; VERMA, Kshitij; JAIN, Verma. Sentence Semantic Similarity Using Dependency Parsing. In: **10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)**. Kanpur, India, 2019.

VUONG, Annalies; NIXON, Tristan; TOWLE, Brendon. A Method for Finding Prerequisites Within a Curriculum. In: **4th International Conference on Educational Data Mining (EDM)**. Eindhoven, Holanda, 2011.

WANG, Shuting; LIU, Lei. Prerequisite concept maps extraction for automaticassessment. In: **25th International Conference Companion on World Wide Web**. Montréal, Canada, 2016.

WANG, Shuting; ORORBIA, Alexander G.; WU, Zhaohui; WILLIAMS, Kyle; LIANG, Chen; PURSEL, Bart; GILES, Clyde L. Using prerequisites to extract concept maps from text books. In: **25th ACM International on Conference on Information and Knowledge Management**. Indianapolis, Estados Unidos, 2016.

YANG, Yiming; LIU, Hanxiao; CARBONELL, Jaime; MA, Wanli. Concept graph learning from educational data. In: **The Eighth ACM International Conference on Web Search and Data Mining (WSDM)**. Shanghai, China, 2015.

ZHOU, Yang; XIAO, Kui. Extracting Prerequisite Relations Among Concepts in Wikipedia. In: **International Joint Conference on Neural Networks (IJCNN)**. Budapeste, Hungria, 2019.

ZHU, Ganggao; IGLESIAS, Carlos A. Exploiting semantic similarity for named

entity disambiguation in knowledge graphs. **Expert Systems with Applications**. n. 101, pp. 8-24, 2018.

APÊNDICE A – SPARQL

A SPARQL (PRUD'HOMMEAUX et. al., 2008), assim como a RDF, foi elaborada pela W3C como uma linguagem padrão para consulta padrão para o RDF.

A consulta SPARQL Pérez et al. (2009), por padrão, é constituída por três partes:

- Padrão de consulta – funcionalidades ao padrão de consulta (escolha do prefixo do dado a ser combinado, partes opcionais, união entre padrões e filtragem de possíveis combinações de valores);
- Modificadores de solução – com o resultado obtido, permite a utilização de operadores (distinct, limit e projections);
- Saída – resultado com vários tipos (construção de um novo RDF, seleção dos valores das variáveis com os padrões de tripla combinados, descrição de recursos).

Como resultado de consultas SPARQL, um conjunto de triplas (padrões) são apresentados, os Basic Graph Patterns, os quais se assemelham às do RDF (PRUD'HOMMEAUX et. al., 2008).

A seguir, são apresentadas as consultas SPARQL desenvolvidas e submetidas na instalação da DBpedia. Elas foram usadas para consultar – na instalação da DBpedia – os conceitos c_i da função indicadora de referência e das funções de peso que fazem parte do cálculo do OP-RefD.

Na composição da fórmula do RefD (2.1), a parte do numerador é composta pela função peso multiplicada por função de indicador.

SPARQL Utilizada na Função Indicadora de Referência:

```
PREFIX :    <http://dbpedia.org/resource/>
PREFIX dbo:    <http://dbpedia.org/ontology/>
PREFIX dbp:    <http://dbpedia.org/property/>
SELECT DISTINCT ?s1 FROM <http://dbpedia.org>
        WHERE  {
        ?s1 ?p1 %s  .
        FILTER (?p1 NOT IN (dbo:wikiPageDisambiguates,
dbp:wikiPageUsesTemplate, dbo:wikiPageRedirects))
        FILTER(STRSTARTS(STR(?s1), 'http://dbpedia.org/resource/'))
        }
```

Onde **%s** é um dos dois conceitos do par conceitual que estiver em avaliação. Por exemplo, avaliando o conceito B, este SPARQL retorna todos os conceitos cis que, no grafo da DBpedia, apontam para o conceito B através de uma propriedade.

SPARQL Utilizada nas Funções de Peso:

```
PREFIX :    <http://dbpedia.org/resource/>
PREFIX dbo:    <http://dbpedia.org/ontology/>
PREFIX dbp:    <http://dbpedia.org/property/>
SELECT  ?p1 ?o1 FROM <http://dbpedia.org>
        WHERE  {
        %s ?p1  ?o1.
        FILTER(?p1 NOT IN (dbo:wikiPageDisambiguates,
dbp:wikiPageUsesTemplate, dbo:wikiPageRedirects))
        FILTER(STRSTARTS(STR(?o1), 'http://dbpedia.org/resource/'))
        }
```

Onde **%s** é um dos dois conceitos do par conceitual que estiver em avaliação. Por exemplo, avaliando o conceito A, este SPARQL retorna todos os conceitos cis do grafo da DBpedia para os quais o conceito A aponta, através de uma propriedade.