UNIVERSIDADE FEDERAL DE ALAGOAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA

WILLIAMS LOURENÇO DE ALCANTARA

Um motor de inferência para relações de identidade em grafos de conhecimento

Maceió-AL

Dezembro de 2019

WILLIAMS LOURENÇO DE ALCANTARA

Um motor de inferência para relações de identidade em grafos de conhecimento

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal de Alagoas.

Orientador: Evandro de Barros Costa

Maceió-AL

Dezembro de 2019

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 - 1767

A347m Alcantara, Williams Lourenço de.
Um motor de inferência para relações de identidade em grafos de conhecimento / Williams Lourenço de Alcantara. – 2020.
54 f. : il.

Orientador: Evandro de Barros Costa.
Dissertação (mestrado em Informática) - Universidade Federal de Alagoas. Instituto de Computação. Maceió, 2019.

Bibliografia: f. 45-51.
Apêndices: f. 52-54

1. Grafos de ligação. 2. Inferência estatística. 3. Web semântica. I. Título.

CDU: 004.423.26

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Programa de Pós-Graduação em Informática – PpgI
Instituto de Computação
Campus A. C. Simões BR 104-Norte Km 14 BL 12 Tabuleiro do Martins
Maceió/AL - Brasil CEP: 57.072-970 | Telefone: (082) 3214-1401

Folha de Aprovação

Williams Lourenço de Alcantara

"Um Motor de Inferência para Relações de Identidade em Grafos de Conhecimento"

Dissertação submetida ao corpo docente do Programa de Pós-Graduação em Informática da Universidade Federal de Alagoas e aprovada em 12 de dezembro de 2019.

Banca Examinadora:

Prof. Dr. Evandro de Barros Costa
UFAL – Programa de Pós-graduação em Informática
Orientador

Prof. Dr. Patrick Henrique da Silva Brito
UFAL – Programa de Pós-graduação em Informática
Examinador Interno

Prof. Dr. Frederico Luiz Gonçalves de Freitas
UFPE – Universidade Federal de Pernambuco
Examinador Externo

Prof. Dr. Mirko Barbosa Perkusich
IFPB – Instituto Federal da Paraíba
Examinador Externo

Aos meus pais Josué Lourenço de Alcantara e Josefa Januário de Alcantara,
por terem vivido este sonho comigo.

AGRADECIMENTOS

Agradeço a Deus.

Agradeço aos meus pais, Josué Lourenço de Alcantara e Josefa Januário de Alcantara, por terem me ensinado a ser persistente e por sempre se esforçarem para me manter firme em busca dos meus objetivos.

À minha esposa Maria Leandra Madeiro de Souza Alcantara, pela compreensão, companheirismo e incentivo nos momentos difíceis.

Ao meu irmão Welton Lourenço de Alcantara, pela parceria e imensa contribuição em nossos projetos.

À minha sogra Luiza Lima de Souza por sempre estar me apoiando.

Aos meus tios Luiz Gomes de Souza e Josete Lourenço de Souza por todo acolhimento e suporte no começo da jornada.

Agradeço ao meu orientador, Evandro de Barros Costa, por todos os conselhos, pela paciência, direcionamentos de pesquisa e os valiosos ensinamentos que tive a oportunidade de receber.

Ao professor Roberio Santos por todos os conselhos e direcionamentos ao longo da pesquisa.

Aos meus grandes amigos Judson Melo Bandeira, Armando Barbosa Sobrinho e Denys Fellipe Souza Rocha que sonham comigo diariamente em busca de novos desafios.

Aos demais familiares e amigos, pela compreensão nos momentos de ausência e por sempre torcerem por mim.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro para realização desta pesquisa.

A todas as pessoas que, de alguma forma, contribuíram para o meu crescimento pessoal, acadêmico e profissional.

“A persistência é o menor caminho do êxito”
(Charles Chaplin)

RESUMO

A crescente demanda por acesso a informações em tempo real tem demandado um alto custo – financeiro e computacional – para a integração de dados devido à ausência de padronização, o que normalmente resulta em problemas durante a modelagem e representação de dados. Grafos de Conhecimento (GC) tem sido um dos mecanismos utilizados para abordar tais problemas. Apesar de fornecerem um modelo estruturado, escalável e compreensível por máquina, a sua criação e manutenção são vulneráveis a erros em razão da dificuldade de raciocínio automático em um grande volume de dados de diferentes domínios – que pode gerar resultados imprecisos, errados ou incompletos – principalmente relacionados à ambiguidade. Normalmente, os problemas causados pela ocorrência de relações ambíguas são derivados da imprecisão ao determinar uma Relação de Identidade (IR) em um domínio, pois os trabalhos existentes na literatura realizam comparação de todos os atributos sem considerar que alguns podem ser mais relevantes. Assim, este trabalho propõe um mecanismo automático para detecção de IR capaz de realizar seleção automática de atributos relevantes de um domínio a partir de análises de entropia e correlação estatística entre os atributos. A solução proposta foi aplicada em 12 conjuntos de dados reais que contém atividades de desenvolvimento de software, sendo que os atributos selecionados automaticamente obtiveram melhor acurácia na detecção de IR do que os atributos relevantes definidos por um especialista do domínio.

Palavras-chave: Relação de Identidade, Grafo de Conhecimento, Inferência, Seleção de Atributos.

ABSTRACT

The growing demand for realtime information access requires high cost – financial and computational – for data integration due to lack of standardization, resulting in problems during modeling and display data. The Knowledge Graphs were used to deal these problems. By providing a structured, scalable and understandable machine model, the creation and maintenance are vulnerable to errors due to automatic reasoning difficulties in large data from different domains – which can produce inaccurate, erroneous or incomplete results – mainly related with ambiguity. The problems are normally caused by ambiguous relationships and by inaccuracy in determining Identity Relations (IR) in a domain. Recent studies compare all attributes without considering that some of them can be more relevant. This work applied an automatic IR detection mechanism which execute an automatic selection of relevant attributes for a domain from entropy analysis and statistical correlation between the attributes. The proposed solution was applied in 12 real datasets that include software development activities. The characters which were automatically selected obtained better IR detection accuracy than the criteria recommended by a domain expert.

Keywords: Identity Relationship, Knowledge Graph, Inference, Attribute Selection.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
FN	Falso Negativo
FP	Falso Positivo
GC	Grafo de Conhecimento
HTTP	Hypertext Transfer Protocol
HTTPS	Hyper Text Transfer Protocol Secure
IBGE	Instituto Brasileiro de Geografia e Estatistica
IR	Relação de Identidade
IRI	Inferência de Relação de Identidade
ISE	Laboratório de Engenharia de Software Inteligente
OSM	Open Street Map
OWL	Ontology Web Language
QP	Questão de Pesquisa
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
SPARQL	SPARQL Protocol and RDF Query Language
SUS	Sistema Único de Saúde
UFCG	Universidade Federal de Campina Grande
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

Recentemente, tem se verificado um progresso significativo na área de inteligência artificial e, particularmente, no contexto de grafo de conhecimento e aprendizagem de máquina, motivando investimentos relevantes tanto da academia, quanto da indústria. De fato, com a popularização da Web, amplia-se cada vez mais o interesse das pessoas em analisar dados para gerar valor e extrair informações que auxiliem na tomada de decisão para realização de tarefas do cotidiano. Particularmente, com o avanço de Internet das Coisas e Computação em Nuvem, diariamente surgem novos dispositivos e aplicações conectadas a Internet que usam diversas fontes de dados de diversos domínios (BOTTA et al., 2016; WANG; KUNG; BYRD, 2018) – transporte, astronomia, saúde, segurança – para fornecer predições e recomendações para seus usuários.

Com a crescente quantidade de dados disponíveis na Web e o surgimento de soluções que auxiliam na organização e padronização desses dados, há a necessidade de conexão de dados em larga escala (BIZER; HEATH; BERNERS-LEE, 2011; HEATH; BIZER, 2011), conectando diversas fontes, aperfeiçoando a capacidade de descoberta de conhecimento, possibilitando que elas consigam raciocinar sobre o seu próprio conhecimento e auxiliem na tomada de decisão. Com isso, os sistemas baseados em conhecimento passam a contar com um ambiente favorável para explorar essa massa de dados e, ainda, minimizar os efeitos de isolamento dos conjuntos de dados na Web, causados pela publicação de dados não-estruturados e sem a reutilização de modelos de representação, dificultando a sua capacidade de conectar-se a outras fontes.

A mencionada diversidade de aplicações torna os conjuntos de dados vulneráveis a baixa capacidade de interoperabilidade causada pelo fato das equipes de desenvolvimento utilizarem diferentes tecnologias, metodologias e processos de desenvolvimento de software que não convergem com a utilização de vocabulários padronizados para a modelagem e representação de dados (BERNERS-LEE et al., 2001). Como estes vocabulários ainda não tem sido adotados em larga escala pela academia e indústria, consequentemente, causa o isolamento de dados na Internet – dificultando o consumo em larga escala.

Em geral, a baixa interoperabilidade limita a capacidade de exploração de um conjunto de dados, pois o torna suscetível à ausência e corrupção de dados (CHEN; ZHANG, 2014), de tal modo, que suscita a importância da utilização de um modelo de representação estruturado, escalável e compreensível por máquina para possibilitar o enriquecimento da base de conhecimento através da integração de dados publicados por diferentes fontes (DONG; SRIVASTAVA, 2013; AUFAURE et al., 2016).

Apesar de ainda não haver consenso acerca de uma definição clara sobre *Grafo de Conhecimento (GC)* (EHRLINGER; WÖSS, 2016), nós assumimos as características compartilhadas por várias definições disponíveis na literatura (LIN et al., 2015; GUAN; SONG; LIAO, 2019;

PAULHEIM, 2017; BLUMAUER, 2014; FÄRBER et al., 2018; PUJARA et al., 2013) para identificar um GC, e a partir delas consideramos que um GC deve, no mínimo, ser capaz de descrever entidades, suas relações e possibilitar a descoberta de conhecimento. Devido à falta de uma definição formal, (PAULHEIM, 2017) apresenta um conjunto mínimo de características que um GC deve ter:

1. descrever entidades do mundo real e suas interrelações, organizadas em um grafo
2. definir possíveis classes e relações de entidades em um esquema
3. permitir interrelacionar entidades potencialmente arbitrárias umas com as outras
4. abranger vários domínios

Ao sugerir que um grafo só pode ser classificado como um GC se possuir uma quantidade relevante dos domínios existentes no mundo, (PAULHEIM, 2017) está limitando o escopo para uma quantidade muito restrita de grafos (tais como DBPedia, FreeBase, Google Vault, Microsoft Sartori, e poucos outros). Apesar de considerarmos relevante a necessidade de ser multi-domínio, esse critério não pode ser fundamental para determinar a existência de um GC, pois não apresenta uma forma clara e objetiva para avaliar se uma determinada quantidade de domínios é suficiente e adequada para tornar um grafo em GC, e principalmente, pelo fato de que qualquer grafo que possua um único domínio pode gerar conhecimento.

De acordo com os princípios de *Dados Conectados* (BIZER; HEATH; BERNERS-LEE, 2011), é possível que um conjunto de dados conecte-se a outras fontes de forma fácil e escalável, independentemente de ser restrito a um único domínio. Assim, tal como um GC, em vez de centralizar todo o conhecimento em um único e grande grafo, é possível aprimorar a descentralização do conhecimento em grafos menores, de forma que possuam uma alta capacidade de integração entre si.

Dessa forma, um GC agrega importantes características que preservam a semântica dos dados de forma estruturada, escalável e independente de domínio. Para exemplificar, vamos considerar um cenário hipotético, no qual, uma pessoa que gosta de praticar exercícios físicos ao ar livre pode estar utilizando um relógio inteligente que pode facilmente monitorar os batimentos cardíacos, quantidade de passos, velocidade e localização. Além disso, conectaria ao telefone celular que poderia lhe sugerir músicas motivacionais – baseado no desempenho atual do corredor – para ajudá-lo a alcançar a sua meta de exercícios, e ainda, fornecer informações sobre fatores externos que podem auxiliá-lo na escolha do melhor momento de voltar para casa, baseado em informações de contexto com as condições climáticas, trânsito, eventos e até mesmo ofertas de produtos e serviços que estão ocorrendo naquela região. Assim, um GC pode ser utilizado para representar e integrar as diversas fontes de dados de vários domínios de forma escalável e prover informações relevantes para auxiliar o usuário na tomada de decisão.

Apesar disso, na construção de um GC nenhuma abordagem pode garantir um resultado perfeito, visto que a construção *manual* é inviável por estar lidando com um grande volume de dados e dificilmente conseguirá alcançar uma cobertura razoável, além de ser necessário reprojetar uma nova solução para cada domínio (HALPIN; MCNEILL, 2013), enquanto a construção *automática* é mais eficiente e possui alta capacidade de escala, no entanto, utiliza métodos heurísticos que são representações aproximadas do mundo real e pode apresentar resultados imprecisos.

Como todas as formas de representação do mundo real são imperfeitas e qualquer imperfeição pode ser fonte de erro (DAVIS; SHROBE; SZOLOVITS, 1993), é comum que as abordagens citadas sejam propensas a erros quando lidam com ambiguidade, condicional, contradição, conhecimento fragmentado, inerte, mal classificado ou incerto (REED; PEASE, 2017).

Um estudo de caso realizado com as bases do DBpedia[1] e Freebase[2] apresenta que 20% das conexões são incorretas (ZAVERI et al., 2013) e aproximadamente 50% das conexões de identidade *owl:sameAs* conectam recursos que possuem algum grau de similaridade mas que não são idênticos (HALPIN et al., 2010). Devido a dimensão dessas duas fontes, com imensa quantidade de dados e diversos domínios, é difícil implementar mecanismos automáticos de inferência com restrições dos diversos domínios, e por isso, tem sido frequente a utilização de abordagens genéricas para identificar a similaridade de recursos.

Para melhorar o desempenho de raciocínio baseado em conhecimento é necessário seguir representações qualitativas para focar apenas em aspectos que são essenciais para realizar uma determinada tarefa, desprezando os não-essenciais (DYLLA et al., 2017). Dessa forma, para inferir uma relação de identidade deve-se considerar que alguns pares propriedade-valor são mais importantes que outros (HU; JIA, 2015; WANG; ENGLEBIENNE; SCHLOBACH, 2008) e que a importância desses pares pode variar de acordo com o contexto (RAAD; PERNELLE; SAÏS, 2017; BEEK; SCHLOBACH; HARMELEN, 2016; RAAD et al., 2018), por isso, é importante manter um equilíbrio entre a generalização e especialização das relações de identidade para minimizar os riscos de imprecisão e fornecer uma solução escalável.

O ciclo de vida de um GC apresenta vários desafios que vem sendo estudados neste campo de pesquisa pela academia e a indústria:

1. **Extração de conhecimento**: a extração de conhecimento lida com a ausência de uma modelagem em comum entre fontes heterogêneas e não estruturadas. Isto torna difícil o processamento automatizado pela máquina devido à subjetividade envolvida nos dados, e tem demandando abordagens manuais, não-supervisionadas ou semi-supervisionadas (PAULHEIM, 2017; WANG et al., 2017).

[1] https://wiki.dbpedia.org/
[2] https://developers.google.com/freebase/

2. **Desambiguação**: requer o gerenciamento da identidade dos recursos para que seja possível identificá-los nas diferentes formas de apresentação. É necessário perceber as especialidades de cada recurso, sem ambiguidade, para que seja possível descobrir novos conhecimentos com o maior nível de precisão.

3. **Evolução do conhecimento**: o GC deve ser flexível o suficiente para adaptar-se as mudanças do mundo real e predizer novos links, pois eventualmente os recursos alteram o seu estado ou comportamento, por exemplo, uma empresa é adquirida por outra ou pode dividir-se em mais de uma (PAULHEIM, 2017; WANG et al., 2017; GOYAL; FERRARA, 2018; CAI; ZHENG; CHANG, 2018).

4. **Escalabilidade**: Com o grande volume de dados – milhões de nós e arestas – em um GC e a crescente demanda por acesso à informação em tempo real, os métodos devem ser capazes de realizar os processos e tarefas em um tempo satisfatório (PAULHEIM, 2017; WANG et al., 2017; GOYAL; FERRARA, 2018; CAI; ZHENG; CHANG, 2018).

1.1 Motivação

Com a crescente demanda por acesso às informações em tempo real que tem sido evidenciada ultimamente, a evolução do conhecimento torna-se imprescindível para garantir a entrega de informações relevantes para as diversas tarefas do cotidiano. Para a integração com outras fontes de dados é fundamental a descoberta de recursos iguais, independentemente da forma na qual estejam identificados. A estratégia de determinar a similaridade de recursos apenas se todas os seus atributos forem iguais, limita a descoberta de conhecimento, visto que, tradicionalmente as diferentes fontes de dados não tem seguido os mesmos padrões de representação, fato que aumenta a probabilidade de haver ambiguidade nos valores dos atributos.

Para ilustrar um cenário de ambiguidade, considere que uma aplicação para recomendação de imóveis utiliza os dados disponibilizados pelo OpenStreetMap[3] (OSM) para apresentar informações básicas sobre cidades, tais como: *nome*, *tradução do nome em diversas línguas*, *população* e *polígono da cidade*. Para auxiliar à tomada de decisão durante a escolha de um imóvel, a aplicação pode ser integrada também com os dados fornecidos pelo Instituto Brasileiro de Geografia e Estatística[4] (IBGE) para agregar informações qualitativas sobre as cidades, tais como: *Índice de Desenvolvimento Humano Municipal*, *Estabelecimentos de Saúde SUS*, *taxa de mortalidade infantil* e *Urbanização de vias públicas*. Apesar destes conjuntos de dados possuírem informações complementares, eles também compartilham os atributos *código do IBGE*, *nome da cidade* e *tamanho populacional* – conforme ilustrado na Figura 1. Ao aplicar um

[3] O OpenStreetMap é desenvolvido por uma comunidade voluntária de mapeadores que contribuem e mantêm atualizados os dados sobre estradas, trilhos, cafés, estações ferroviárias e muito mais por todo o mundo. Disponível em <https://www.openstreetmap.org/>

[4] O Instituto Brasileiro de Geografia e Estatística - IBGE se constitui no principal provedor de dados e informações do País, que atendem às necessidades dos mais diversos segmentos da sociedade civil, bem como dos órgãos das esferas governamentais federal, estadual e municipal. Disponível em <https://www.ibge.gov.br>

mecanismo automático para realizar inferências de relações de identidade entre as cidades serão comparados todos os seus atributos, logo, só existirá uma relação de identidade se todos os pares atributo-valor forem iguais. Como a quantidade populacional nos dois conjuntos de dados são diferentes, conforme Figura 1, o motor de inferência não conseguirá identificar que a cidade de Maceió originada no OSM possui uma equivalente na base do IBGE, apesar do código IBGE ser igual em ambas. Dessa forma, as informações qualitativas desta cidade estarão ausentes na aplicação devido à ambiguidade causada pelo conflito de valores para um atributo.

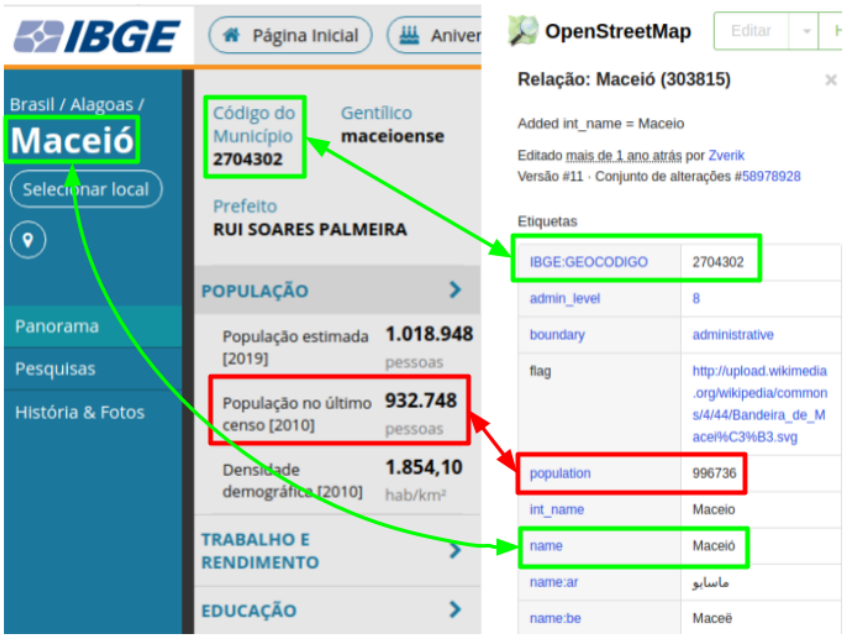

Figura 1 – Ausência de relação de identidade

Fonte: IBGE <https://cidades.ibge.gov.br/brasil/al/maceio/panorama> e OSM <https://www.openstreetmap.org/relation/303815>

As diversas formas de ambiguidade pode causar transtornos para muitas pessoas na execução de tarefas do cotidiano. Por exemplo, uma pessoa que desejava viajar para a cidade Granada situada na Espanha e foi, involuntariamente, para uma homônima no Caribe. Nesse caso, a situação ambígua foi causada pelo fato da pessoa que comprou a passagem não conseguiu distinguir as diferentes cidades, conforme Figura 2.

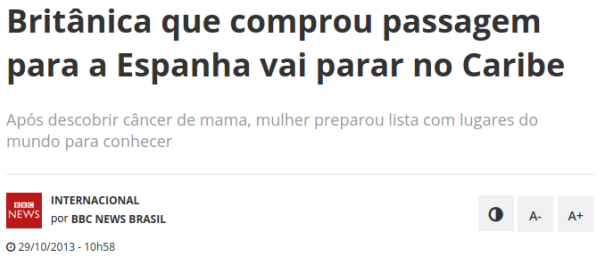

Figura 2 – Cidades homônimas

Fonte: BBC News Brasil (<http://tiny.cc/t8vtgz>)

Considerando o raciocínio dedutivo da lógica de descrição, sob a perspectiva da psicologia (REED; PEASE, 2017; JOHNSON-LAIRD, 1999), um GC pode alcançar relações simétricas e transitivas que são capazes de derivar novos fatos na base de conhecimento (BAADER; HORROCKS; SATTLER, 2008; HORROCKS; SATTLER; TOBIES, 1999), de tal modo que um fato novo pode gerar um efeito cascata que vai derivar outros novos fatos. Assim, quando uma dedução é baseada em premissas erradas, consequentemente, irá gerar um fato também errado, por dedução lógica. Nesse caso, o efeito cascata acarretará em uma propagação de derivações inválidas que pode causar grandes prejuízos à confiabilidade das aplicações.

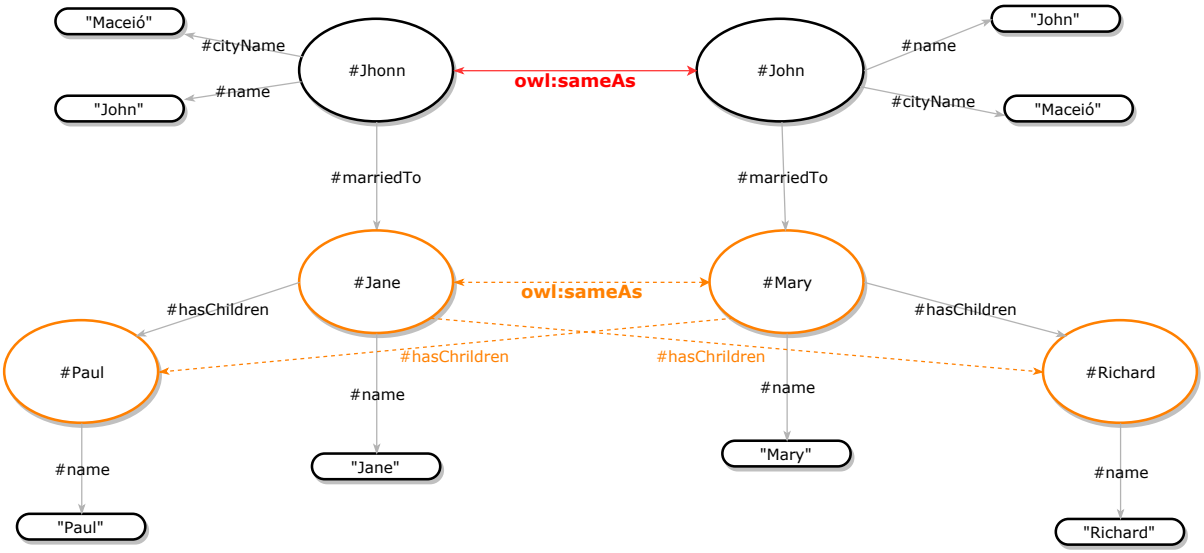

Figura 3 – Efeito cascata da inferência dedutiva inválida

Fonte: Autor (2019)

Para exemplificar, no cenário hipotético apresentado na Figura 3, considere que a relação de identidade (*owl:sameAs*) entre as instâncias *#John* e *#Jhonn*, que possuem identificadores diferentes, foi atribuída após utilizar os atributos *#name* e *#cityName* para efetuar a medida de similaridade comparando os pares propriedade-valor. Consequentemente pela característica funcional do predicado *#marriedTo* novas derivações são realizadas atribuindo uma relação de similaridade entre as instâncias *#Jane* e *#Mary*, e, por consequência lógica definindo que *#Paul* e *#Richard* são filhos da mesma mãe. Agora se considerar que os atributos selecionados são insuficientes para determinar a similaridade, visto que pessoas diferentes podem ter nomes iguais e morarem na mesma cidade, ocorre o problema causado por uma dedução realizada a partir de premissas inválidas.

Apesar de existirem vários estudos que propõem mecanismos de integração de dados, a correção de conexões entre instâncias ainda não é suficientemente abordada na literatura (PARIS, 2018), fazendo-se necessário aprofundar os estudos acerca de *relações de identidade* para prover mecanismos de inferência mais assertivos. As abordagens existentes são insensíveis à restrições de domínio, normalmente, elas consideram pares propriedade-valor iguais para

determinar a similaridade (LI et al., 2009; ARAUJO et al., 2012), sem avaliar se os atributos são relevantes para inferir a identidade de um recurso, como os atributos *#name* e *#cityName* citados anteriormente.

No contexto de evolução em GC, uma IR pode ser determinada pela análise semântica, com o suporte de RDFS/OWL. No entanto, mesmo com a análise semântica bem resolvida, a ambiguidade ainda pode persistir por causa da influência da seleção de atributos irrelevantes que ocorre ao não considerar que alguns atributos são mais relevantes no domínio.

O escopo desta pesquisa é investigar a influência da análise sintática na seleção automática de atributos relevantes para determinar uma IR, por isso, a análise semântica não é objeto de estudo desta pesquisa. Além disso, ao realizar inferências a partir da análise sintática é utilizado apenas *inferência estatística*.

1.2 Problema

A definição do significado de recursos é uma tarefa subjetiva que pode gerar ambiguidade na compreensão de seus relacionamentos na base de conhecimento, pois pode ser influenciada pelo contexto e pela percepção do observador (HALPIN; MCNEILL, 2013). A comunidade tem buscado solucionar a desambiguação durante a descoberta de recursos iguais em uma base de conhecimento apontando soluções para o conflito de objetos (LIU et al., 2019). No entanto, antes disso é necessário resolver a ambiguidade na seleção de atributos irrelevantes para determinar a identidade de um recurso. Existem várias abordagens que dependem de grande quantidade de dados anotados *manualmente* para se adaptarem a um domínio diferente na tentativa de contribuir para solucionar a desambiguação (RIZZO; ERP; TRONCY, 2014). Quando um recurso é analisado isoladamente pode gerar confusão acerca do que ele pode representar. Por isso, é necessário reunir as informações relevantes para minimizar os riscos de ambiguidade.

1.3 Questões de pesquisa e Objetivos

O objetivo deste trabalho foi desenvolver uma solução que auxilie na seleção automática de atributos para que possa gerar contribuições nos estudos acerca de desambiguação e evolução de conhecimento. Em geral, deseja-se melhorar a capacidade de descoberta de relações de identidade. Para isso, a seguinte questão de pesquisa foi definida:

QP: Como pode ser melhorado o desempenho de inferência de relações de identidade?

Com a resposta para esta questão de pesquisa, pode-se alcançar os seguintes objetivos específicos:

1. *Seleção automática de atributos*: prover um mecanismo automático para medir a relevância de cada atributo em um domínio, mensurando o seu potencial de ser um identificador de

um registro do conjunto de dados.

2. *Redução da dimensionalidade*: reduzir o espaço de processamento de dados para a realização de uma inferência através da análise de contexto para identificar os atributos que estão fora do escopo da aplicação.

1.4 Relevância da pesquisa

A todo momento as pessoas buscam informações para aprofundar o conhecimento sobre tópicos de interesse, com isso, a conexão com outras bases de dados é fundamental para o bom desempenho das atividades no cotidiano através do enriquecimento da base de conhecimento. O potencial limitado de perceber recursos iguais a partir de diferentes fontes de forma automática pode gerar grandes prejuízos – financeiros e sociais – para a sociedade e organizações.

Estima-se que o desvio de dinheiro público no Brasil desviou, no mínimo, 40 bilhões de dólares[5] nos últimos anos. Em investigações criminais contra a corrupção, comumente, ocorre o cruzamento de informações de vários conjuntos de dados em busca de informações complementares sobre uma mesma pessoa ou entidade, as quais, estão contidas em grandes volumes de dados e podem estar identificadas de diferentes formas – dificultando a percepção de similaridade de forma automática. Com isso, o combate a corrupção torna-se limitado devido à baixa capacidade de percepção de recursos similares, o que poderia facilitar o provimento de informações detalhadas que contribuiriam para a elucidação de crimes.

1.5 Estrutura

Esta dissertação está organizada da seguinte forma: Capítulo 2 apresenta o embasamento teórico de alguns conceitos envolvidos em Grafo de Conhecimento, e, uma visão geral da estrutura de Sistemas baseados em conhecimento e em Grafos de Conhecimento: métodos, escopo, arquitetura, ciclo de vida e problemas comuns. Capítulo 3 discute o funcionamento dos principais trabalhos relacionados e traz reflexões sobre as estratégias adotadas para alcançar a descoberta automática de relação de identidade. Capítulo 4 apresenta a solução desenvolvida neste trabalho, detalhando a metodologia e avaliação do estudo de caso. Capítulo 5 apresenta os resultados, discussões e ameaças à validade. Capítulo 6 apresenta conclusões e trabalhos futuros.

[5] Em cinco anos à frente da Lava Jato, o procurador prendeu corruptos e recuperou R$ 12 bilhões desviados da Petrobras <https://istoe.com.br/em-cinco-anos-a-frente-da-lava-jato-o-procurador-prendeu-corruptos-e-recuperou-r-12-bilhoes-desviados-da-petrobras/?fbclid=IwAR0yEOaL0Xhqhvh-nZ72yQW8mkybDMpiwiy42-YJtvWCrsfanEwYZEKtHOw>

2 FUNDAMENTAÇÃO

Esta seção apresenta os conceitos mais relevantes nos quais grafos de conhecimento são baseados. Serão fornecidas descrições gerais sobre gerenciamento e representação do conhecimento, bem como, o embasamento necessário para compreendê-las.

2.1 Hierarquia informacional

Em uma comunicação entre pessoas ocorre constantemente a troca de informações derivadas de observações sob a perspectiva do interlocutor. Considerando que há uma grande quantidade de dados e informações sendo disseminadas em larga escala, que são capazes de influenciar na tomada de decisões, surge a necessidade de gerenciamento do conhecimento aplicando métodos e princípios para qualificar a informação e transformá-la em conhecimento verdadeiro para alcançar valores confiáveis que, naquele momento, representam uma boa solução para uma determinada necessidade (ACKOFF, 1989; BELLINGER; CASTRO; MILLS, 2004).

Em meio a esta grande massa de dados surge um perigo evidente quando informações falsas são transmitidas com o objetivo de manipular as decisões tomadas pelas pessoas, fazendo-as convergirem para um conhecimento falso (BU; XIA; WANG, 2013) que não condiz com a realidade. Nos últimos anos, este tipo de informação tem levado algumas nações a refletirem sobre suas crenças e princípios, influenciando decisões políticas (BOVET; MAKSE, 2019; GRAHAM, 2017), fazendo-as disseminarem conteúdo sem qualquer tipo de checagem sobre a sua veracidade. Isso causa um efeito cascata, criando uma nuvem de informações sem checagem que são cada vez mais disseminadas e utilizadas para gerar interpretações equivocadas sobre a realidade.

De acordo com (ACKOFF, 1989) o conteúdo do conhecimento pode ser organizado nas seguintes categorias:

1. *Dado* - são registros de observações do mundo real que não trazem significado além de sua existência.
2. *Informação* - analisa e apresenta uma interpretação baseada nos dados.
3. *Conhecimento* - consiste de uma coleção de informações que refletem uma experiência, ele é baseado em casos que já ocorreram.
4. *Compreensão* - ocorre quando um novo conhecimento é gerado a partir do conhecimento adquirido previamente.
5. *Sabedoria* - está associada a julgamentos a cerca da eficácia de uma ação tomada, uma avaliação que é dependente do julgador - baseado na sua experiência e compreensão do objeto julgado.

A compreensão e o conhecimento referem-se a "aprendizagem"e "memorização", respectivamente. A compreensão possibilita aprender com base no que é previamente conhecido (conhecimento adquirido), e ainda, determina novas ações úteis capazes de gerar novos conhecimentos (BELLINGER; CASTRO; MILLS, 2004).

As quatro primeiras categorias estão associadas a eficiência, enquanto a última diz respeito a eficácia dos valores gerados. A avaliação da eficiência é baseada na lógica e pode ser reproduzida por computadores, independentemente do julgador. No entanto, a avaliação da eficácia traz consigo uma relação pessoal, que pode divergir de acordo com a perspectiva do julgador (ACKOFF, 1989).

Para alcançar altos níveis da hierarquia informacional é importante adotar boas práticas para representação de conhecimento em larga escala e com expressividade semântica.

2.2 Representação de conhecimento

É um campo da Inteligência Artificial cujo objetivo é simular a inteligência humana em uma linguagem de computador, de modo que o computador seja capaz de resolver problemas do mundo real (GROSAN; ABRAHAM, 2011; BARAL; GELFOND, 1994; HOEKSTRA, 2009).

Nessa evidente necessidade de compreender o comportamento dos humanos, o especialista do domínio desempenha um papel determinante para alcançar uma representação de conhecimento bem sucedida, é necessário considerar restrições que possuem algum grau de incerteza ou incompletude, evitar ambiguidade, descriminar entidades de forma estruturada e precisa para que seja possível projetar inferências, responder questões, atualizar a base de conhecimento e determinar o comportamento desejado do programa (GUARINO, 1995). Apesar disso, o conhecimento adquirido pelo especialista pode ser limitado pela experiência e subjetividade sob a perspectiva do observador (LEAKE; MAGUITMAN; REICHHERZER, 2014), além da dificuldade de lidar manualmente com grandes bases de conhecimento (STERCKX et al., 2016). Assim cada especialista pode ter um entendimento diferente acerca do domínio.

Com o avanço de técnicas de aprendizagem de máquina é possível minimizar a subjetividade e ao mesmo tempo basear-se na experiência através da construção automática de bases de conhecimento, em larga escala, para obter *insights* e compreender o comportamento dos usuários a partir de casos concretos do mundo real.

Uma representação de conhecimento deve ser composta pelos seguintes cinco distintos papéis (DAVIS; SHROBE; SZOLOVITS, 1993):

1. *substituto* - um substituto de um objeto do mundo real, que habilita uma entidade determinar consequências por pensar em vez de agir, isto é, raciocinar sobre o mundo em vez de agir nele.

2. *compromisso ontológico* - seleção de restrições relevantes do domínio, como elas realmente são no mundo real, para alcançar raciocínios que reflitam a realidade.

3. *raciocínio inteligente* - derivar uma conclusão a partir de premissas corretas. O comportamento inteligente pode ser influenciado pelo grau de formalidade dos vocabulários.

4. *computação eficiente* - capacidade de organizar informações para facilitar a execução de inferências e descobrir novos conhecimentos.

5. *expressão humana* - linguagem utilizada para expressar coisas sobre o mundo.

Estes papéis, em conjunto, representam um framework que pode ser aplicado em qualquer forma de representação do conhecimento e é capaz de preservar a essência do conhecimento existente no mundo real para o mundo computacional. Apesar de uma representação não ser capaz de cobrir todo o domínio, ela precisa fazer uma seleção das características relevantes e fixar-se em uma perspectiva (HOEKSTRA, 2009) para garantir uma aproximação ao mundo real. Assim, estes princípios de representação de conhecimento fornecem um meio mais fácil para expressar relações semânticas durante o desenvolvimento de aplicações.

2.3 Web Semântica

Atualmente a *World Wide Web (Web)* possui uma grande quantidade de dados em formatos não-estruturados ou semi-estruturados, principalmente na linguagem *Hyper Text Markup Language (HTML)*. Estes formatos são úteis para fornecer informações legíveis para humanos (BERNERS-LEE et al., 2001). Eles podem ler estas páginas de conteúdo, compreender o significado de seus objetos e produzir novas interpretações. No entanto, estes formatos não facilitam o processamento automatizado para que os computadores possam identificar o significado associado a cada recurso da página.

Pela natureza descentralizada de geração de conteúdo na Web, associada com a ausência de padronizações na forma de representação, atualmente, cria-se grandes silos de dados isolados que limitam a sua capacidade de consumo e agregação de forma automatizada. Para organizá-las, tornando-as bem estruturadas, facilitando a compreensão por máquinas e para que as pessoas construam informação de forma colaborativa, a Web Semântica propõe uma extensão da Web para criar um banco de dados global (BERNERS-LEE et al., 2001).

A Web Semântica tem o objetivo de prover um mecanismo de organização e representação dos dados na Web através de um banco de dados conectado globalmente para facilitar o processamento, reúso e integração de vários conjuntos de dados de forma automatizada pelas máquinas. Dessa forma, fornecerá um mecanismo para estruturação do conteúdo, criando um ambiente onde agentes de software naveguem facilmente entre as páginas para que possam facilmente executar tarefas sofisticadas para os usuários (BERNERS-LEE et al., 2001).

Em termos práticos a Web Semântica é implementada por um conjuntos de tecnologias padronizadas pela *World Wide Web Consortium (W3C)*, dentre as quais, destacam-se:

- *RDF*: O Resource Description Framework (RDF) é um framework para a representação de fatos simples no formato de triplas (<sujeito, predicado, objeto>). Ele permite representar informações sobre os recursos na forma de grafos, sendo que o sujeito refere-se um recurso que será descrito, o predicado representa uma propriedade de descrição do sujeito, e o objeto representa um valor para o predicado que está associado a um sujeito. Além disso, as três partes da tripla utilizam URI para identificar unicamente os recursos, e, apenas o objeto aceita também valores de tipos de dados primitivos, tais como: *inteiro, boolean, string, data, etc*.

- *RDFS*: O Resource Description Framework Schema (RDFS) apresenta uma extensão semântica do RDF e é responsável por prover um conjunto de recursos interrelacionados. Ele possibilita a representação de hierarquia de classes, bem como características das relações domínio e alcance dos predicados.

- *OWL*: Ontology Web Language (OWL) estende o RDFS adicionando maior nível de expressividade na representação dos fatos na base de conhecimento. Ele possibilita a representação de restrições adicionais – *funcional, transitiva, simetria, disjunção, inversa, cardinalidade*.

Após este esforço para organização do conhecimento na Web com a construção de um banco de dados global, é necessário tornar os dados navegáveis, conectando os diferentes conjuntos de dados através da similaridade de recursos.

2.4 Dados Conectados

Dados conectados surgiu a partir da necessidade de interoperar dados não-estruturados na Web. Com o objetivo de orientar as organizações a disponibilizarem os dados em formatos estruturados e compreensíveis por máquina foi proposto um conjunto de boas práticas para publicação e conexão de dados na Web (BIZER; HEATH; BERNERS-LEE, 2011) baseado em padrões internacionais do World Wide Web Consortium - W3C[1], permitindo o estabelecimento de uma rede de dados de diferentes fontes que se conectam e auto enriquecem (HEATH, 2011).

Para (BERNERS-LEE, 2006), a Web Semântica não refere-se a simplesmente publicar dados na Web, mas deveria conectá-los a outras fontes para possibilitar que as pessoas encontrem informação útil a partir de dados relacionados. Com isso, ele propôs os seguintes quatro princípios que implementam a Web de dados conectados:

[1] https://www.w3.org

1. Deve ser usado o padrão Uniform Resource Identifier (URI) (BERNERS-LEE; FIELDING; MASINTER, 2004) para identificar as coisas;
2. Devem ser usadas URIs com o protocolo HyperText Transfer Protocol (HTTP)(FIELDING et al., 1999) para que os usuários possam localizar estes nomes;
3. Quando a URI for encontrada, ela deve prover informação útil, usando padrão Resource Description Framework (RDF) (KLYNE; CARROLL, 2006) ou SPARQL Protocol and RDF Query Language (SPARQL) (GROUP et al., 2013);
4. As URIs devem incluir hiperlinks para outras URIs, para que os usuários possam descobrir novos recursos que se relacionem à URI que estejam buscando.

A navegação em dados conectados pode ser vista como uma navegação em grafos, considerando o modelo RDF, os nós são rotulados por sujeitos ou objetos e os arcos por predicados (BERNERS-LEE, 2006).

Existem duas tecnologias que são fundamentais para garantir a viabilidade de Dados Conectados: URI e HTTP. As URIs fornecem um meio genérico para identificar recursos e o HTTP possibilita de maneira simples que os recursos sejam encontrados na Web. Além dessas, o Resource Description Framework (RDF) também torna-se importante por fornecer um modelo de dados baseado em grafo para representar coisas existentes no mundo real. Um modelo RDF é composto por um conjunto de triplas (sujeito, predicado, objeto). O sujeito é definido por uma URI pois identifica um recurso, o objeto pode assumir uma URI ou um literal, e o predicado assume uma URI que indica uma relação entre o sujeito e o objeto. Por fim, as conexões entre recursos idênticos são realizadas com a propriedade *owl:sameAs* que indica que dois recursos de fontes diferentes são equivalentes (BIZER; HEATH; BERNERS-LEE, 2011).

É importante destacar que a conexão de dados de diferentes fontes deve considerar as características dos contextos envolvidos nos recursos similares.

2.5 Sensibilidade ao contexto

O contexto representa um conjunto de propriedades que descrevem recursos (entidades) e auxiliam na percepção do estado destes recursos para fornecer informações úteis a medida que um novo evento possa surgir e alterá-los. Para (DEY, 2001), contexto é qualquer informação que pode ser usada para caracterizar a situação de uma entidade. Se uma parte da informação pode ser usada para caracterizar a situação de um participante em uma interação, então esta informação é o contexto. Existem várias aplicações utilizadas em diversos domínios – iluminação, nível de ruído, conectividade de rede, custos de comunicação, largura de banda de comunicação e até mesmo a situação social – que fornecem informações relevantes para o usuário através da sensibilidade ao contexto, de tal modo, que a relevância depende da tarefa executada por este usuário (SCHILIT et al., 1994).

De acordo com (SCHILIT et al., 1994), a computação sensível ao contexto "adapta-se de acordo com a sua localização, a coleção de pessoas e objetos próximos, bem como as mudanças desses objetos ao longo do tempo". As aplicações deste tipo podem possuir as seguintes características:

1. *apresentação* de informações e serviços para um usuário
2. *execução* automática de um serviço para um usuário
3. *marcação* de contexto com informações para suportar recuperação posterior

Os conflitos ou ambiguidade em contextos é um problema frequente em sistemas baseados em conhecimento e podem causar prejuízos de aprendizagem, visto que pode interpretar contextos de forma equivocada.

2.6 Grafo de conhecimento e Sistema baseado em conhecimento

Em um sistema baseado em conhecimento é importante que consiga desenvolver habilidades de aprendizagem, raciocinando sobre o seu próprio conhecimento e refinando-o para adequar aos seus objetivos particulares (POOLE; MACKWORTH, 2010; CHEIN; MUGNIER, 2008). De forma similar, um sistema baseado em grafo de conhecimento possui as mesmas características e complementa com a alta capacidade de escalabilidade proporcionada pelo modelo de representação baseada em grafo. O termo *Grafo de Conhecimento* surgiu em 2012[2] quando foi utilizado para representar uma coleção de informações, interconectadas, sobre o mundo real (BORDES et al., 2014). O seu objetivo é descrever qualquer *"coisa"* existente no mundo.

2.6.1 Arquitetura

Esta seção tem o objetivo de apresentar uma visão geral da arquitetura de aplicações de grafos de conhecimento baseadas em agentes inteligentes. Normalmente um agente é composto basicamente por quatro componentes, conforme ilustrado na Figura 4, são eles: *Grafo de conhecimento*, *Execução de consultas*, *Motor de inferências* e *Motor de aprendizagem*.

O componente *Grafo de conhecimento* é responsável por armazenar os dados em um modelo de representação em grafo que podem ser criados em tempo real ou até mesmo importar os conjuntos de triplas de aplicações existentes. Além disso, aplica um motor de inferência capaz de analisar a semântica dos relacionamentos com o objetivo de identificar fatos novos.

Execução de consultas provê um meio de acesso aos dados, abstraindo operações (criação, recuperação, atualização e remoção de dados) específicas das aplicações.

Motor de aprendizagem é responsável por descobrir fatos novos relevantes para o domínio de aplicação. Para isso, ele atua a partir da percepção do contexto envolvido no interesse de

[2] https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.html

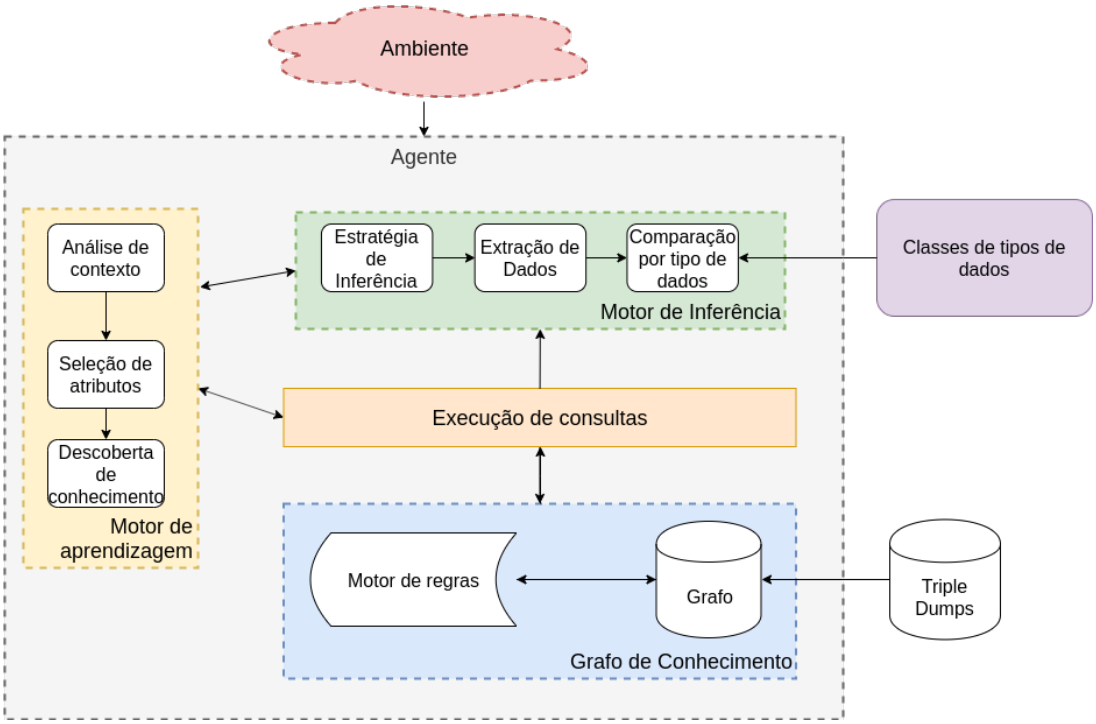

Figura 4 – Arquitetura de grafo de conhecimento

Fonte: Adaptado pelo autor (2019)

entrada do usuário (*Análise de contexto*), em seguida seleciona as características mais relevantes no contexto (*Seleção de Atributos*), por fim, aplica mecanismos de inferência de novos fatos para agregar à base de conhecimento (*Descoberta de conhecimento*).

Motor de inferência é responsável por executar as estratégias de inferência para auxiliar à descoberta de conhecimento. Ela fornece componentes de abstração para facilitar e flexibilizar a aplicação em diversos domínios de forma que seja capaz de suportar diferentes restrições específicas de cada aplicação. Para isso, o componente *Estratégia de inferência* contém detalhes de representação dos modelos de inferência baseado nas características selecionadas. O *Extração de tipos de dados* e *Comparação de tipos de dados* são responsáveis por aplicar as estratégias de comparação das características a partir das classes dos seus tipos de dados (string, boolean, geolocation, image, video, e outros).

2.6.2 O fluxo do grafo de conhecimento

A representação de conhecimento baseada em grafo possibilita uma navegação mais intuitiva de grandes bancos de dados devido a sua capacidade de preservar as características dos recursos. Um GC é composto pelos seguintes três componentes: *(i) entidades* - são os vértices do grafo, representam entidades do mundo real; *(ii) atributos* - são características das entidades, como o tipo; e *(iii) relações* - rotulam as arestas do grafo, representam uma relação entre entidades. Com estes componentes, entidades e atributos não podem possuir ambiguidade e as arestas são identificadas com um conjunto limitado de relações.

Para (PAULHEIM, 2017), um grafo só pode ser classificado como um GC se possuir uma quantidade relevante dos domínios existentes no mundo. Para ele, um conjunto de dados como GeoNames[3] não é um GC porque é restrito a um único domínio. No entanto, é importante ressaltar que esta afirmação foi apresentada de forma imprecisa e não traz referencial teórico e científico para justificá-la. Apesar de considerarmos relevante a necessidade de ser multi-domínio, esse ponto não pode ser fundamental para determinar a existência de um GC, pois não apresenta uma forma clara e objetiva para avaliar se uma determinada quantidade de domínios é suficiente e adequada para tornar um grafo em GC. Seguindo os princípios de *Dados Conectados* (BIZER; HEATH; BERNERS-LEE, 2011), que facilitam a publicação e compartilhamento de dados, é possível que um conjunto de dados conecte-se a outras fontes de forma fácil e escalável, independentemente de ser restrito a um único domínio. Assim, em vez de centralizar todo o conhecimento em um único e grande grafo, é possível aprimorar a descentralização do conhecimento em grafos menores, de forma que possuam uma alta capacidade de integração com outras fontes. Tomando como exemplo o GeoNames, ele segue boas práticas de publicação de dados na Web (SCHMACHTENBERG; BIZER; PAULHEIM, 2014) que possibilitam que ele possa ser integrado a outros conjuntos de dados e recursos da Web de forma fácil, tais como (AUER et al., 2007; REBELE et al., 2016; TRONCY; MALOCHA; FIALHO, 2010; HASSANZADEH; CONSENS, 2009). Dessa forma, não é suficiente e adequado afirmar que GeoNames não pode ser considerado um GC porque possui um único domínio.

Para facilitar a integração entre Grafos de Conhecimento é importante identificar atributos capazes de identificar unicamente os recursos do domínio, para isso é necessário quantificar o nível de relevância de atributos com a análise de variabilidade dos valores baseado no cálculo de entropia.

2.7 Entropia

A identificação de agrupamentos de dados durante o processamento e análise de dados pode auxiliar a identificar registros de dados que possuem comportamentos parecidos. Para calcular a capacidade de agrupamento dos atributos de um conjunto de dados é necessário avaliar a variabilidade dos valores de atributos através da medida de entropia.

A partir de um modelo probabilístico, um preditor que sempre gera o próximo caractere de uma sequência sendo X, terá entropia mínima, porque não há variabilidade. Se o próximo caractere for sempre diferente de todos anteriores, então será atribuído o valor máximo para a entropia.

Inspirado na definição de entropia originária da Termodinâmica, (SHANNON, 1948) define a entropia em Teoria da Informação como uma medida de incerteza de informações

[3] https://www.geonames.org

relevantes em uma sequência (ŠESTÁK; MAREŠ; HUBÍK, 2010). Para ele, a entropia mede a quantidade de informação aleatória em uma variável, conforme Equação 2.1.

$$H(X) = -\sum_{i=0}^{N-1} p_i log_2(p_i) \tag{2.1}$$

Para avaliar o desempenho da abordagem de classificação que utiliza entropia para selecionar atributos relevantes e identificar recursos similares, podem ser utilizadas métricas de avaliação para medir a capacidade e confiabilidade ao realizar classificações positivas e negativas.

2.8 Métricas de avaliação

Para avaliar a confiabilidade das inferências realizadas por um software podem ser utilizadas as métricas *precision*, *recall*, *F-measure*, *informedness* e *markedness* (POWERS, 2011). Elas avaliam a capacidade do software perceber a presença ou ausência de determinadas situações e fornecer resultados corretos.

O cálculo destas métricas envolvem quatro categorias básicas, descritas a seguir:

- *Verdadeiro Positivo (VP)*: registros classificados corretamente como positivos.
- *Falso Negativo (FN)*: registros que foram erroneamente classificados como negativos.
- *Falso Positivo (FP)*: registros que foram erroneamente classificados como positivos. Apesar de terem sido classificados como positivos, eles são negativos.
- *Verdadeiro Negativo (VN)*: registros classificados corretamente como falsos.

Para avaliar um algoritmo de classificação é necessário observar o seu desempenho ao detectar a existência e ausência do objeto de estudo. Para exemplificar, considere um cenário hipotético de detecção de uma doença em humanos. Neste cenário, tão importante quanto acertar que uma pessoa tem câncer, é ter também uma boa capacidade de acerto ao inferir que uma pessoa não tem câncer. Caso o desempenho para uma dessas situações seja baixo, afeta a credibilidade e a confiança acerca das classificações.

A avaliação de identificações positivas pode ser realizada pelas métricas *precision*, *recall* e *F-measure*. A *precision* (Equação 2.2) avalia a capacidade de perceber que existem objetos verdadeiros através da proporção de identificações positivas que foram realizadas corretamente.

$$precision = \frac{VP}{VP + FP} \tag{2.2}$$

A *recall* (Equação 2.3) refere-se a confiabilidade ao classificar um objeto como verdadeiro, analisando a proporção de positivos reais que foram identificados corretamente.

$$recall = \frac{VP}{VP + FN} \tag{2.3}$$

A *F-measure* (Equação 2.4) baseia-se em *precision* e *recall*, na qual atinge seu valor máximo em 1 quando *precision* e *recall* são perfeitas.

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \tag{2.4}$$

Informedness (Equação 2.5) é uma medida de como o sistema está informado sobre pontos positivos e negativos.

$$Informedness = \frac{VP}{VP + FN} + \frac{VN}{FP + VN} - 1 \tag{2.5}$$

Markedness (Equação 2.6) refere-se a confiabilidade das previsões positivas e negativas realizadas pelo sistema.

$$Markedness = \frac{VP}{VP + FP} + \frac{VN}{FN + VN} - 1 \tag{2.6}$$

3 TRABALHOS RELACIONADOS

Este capítulo apresenta uma revisão da literatura e as principais características dos trabalhos existentes que são mais similares à solução proposta nesta pesquisa.

3.1 Revisão da literatura

Nesta seção apresenta-se um detalhamento do processo realizado na revisão da literatura. Para auxiliar na seleção de publicações relevantes para o tópico de estudo, foram definidos critérios de inclusão e exclusão. Estes critérios contribuem para minimizar a possibilidade de inclusão de publicações que não possuem evidência direta com o objetivo desta pesquisa. Assim, os trabalhos que atenderem a todos os critérios de inclusão serão mantidos, caso contrário, estarão fora do escopo e serão ignorados. Os critérios de inclusão são listados a seguir:

- Periódico revisado por pares
- Publicado no período entre janeiro de 2012 e setembro de 2019
- Estudo que propõe uma estratégia para inferência de relação de identidade

O escopo desta revisão cobre soluções que auxiliam na detecção e validação de relações de identidade em GC em qualquer domínio de aplicação. Dessa forma, os trabalhos que não apresentarem contribuições para este tópico de pesquisa estão fora do escopo. Para isso foram definidos os seguintes critérios de exclusão:

- Duplicados
- Não escritos em inglês
- Não propõe uma solução para inferência de relação de identidade (aplicação, metodologia, survey)
- Solução projetada para uma língua específica
- Atas de congressos
- Resenhas
- Capítulo de livro
- Livro
- Contribuição principal do estudo está fora da Ciência da Computação

A busca foi realizada com o uso de palavras-chave que aparecem com frequência em publicações científicas relacionadas ao tópico de pesquisa deste trabalho. O Listing 3.1 apresenta a estrutura lógica dos termos consultados no portal de periódicos da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)[1], resultando em 859 publicações obtidas a partir das seguintes bibliotecas digitais: Scopus[2] (Elsevier), Science Citation Index Expanded[3] (Web of Science), GALE OneFile[4], Advanced Technologies & Aerospace Database[5], ScienceDirect Journals[6], Elsevier[7], MEDLINE/PubMed[8], Springer[9], Social Sciences Citation Index[10] (Web of Science), SpringerLink[11] e Emerald Insight[12].

```
("knowledge graph"
  OR "knowledge base"
  OR "linked data"
  OR "ontology")
 AND ("owl:sameAs"
  OR "identity link"
  OR "identity relation")
 AND ("align"
  OR "match"
  OR "inference")
```

Listing 3.1 – Termos para busca de artigos

A extração das características dos trabalhos foi realizada em duas etapas. Primeiro foram aplicados os filtros automáticos fornecidos pela própria ferramenta de busca (*tipo de revisão do periódico, idioma, data e tipo da publicação*), e por fim, os demais dados (*palavras-chave, título, resumo e contribuição da pesquisa*) foram obtidos de forma manual a partir da leitura completa das publicações. Após a aplicação destes filtros restaram 30 publicações selecionadas, conforme síntese apresentada na Tabela 1.

Para auxiliar na seleção das publicações mais similares foi realizada uma avaliação de qualidade a partir das métricas apresentadas na Tabela 2.

Os critérios C1, C2, C3 e C4 utilizam uma escala binária (Sim/Não). Para cada publicação, se a resposta para um destes critérios for *Sim*, então recebe 1 ponto para o respectivo critério, caso contrário, recebe 0 ponto. Além disso, também foi avaliado a qualidade da revista da publicação através do Fator de Impacto no ano 2018 (ano de medição mais recente). Como este índice de qualidade da revista não é normalizado, nós utilizamos a Equação 3.1 para criarmos um índice

[1] Disponível em <https://www.periodicos.capes.gov.br>
[2] Disponível em <https://www.scopus.com>
[3] Disponível em <http://mjl.clarivate.com/cgi-bin/jrnlst/jloptions.cgi?PC=D>
[4] Disponível em <https://www.gale.com/intl/c/general-science-collection>
[5] Disponível em <https://www.proquest.com/products-services/adv_tech_aero.html>
[6] Disponível em <https://www.sciencedirect.com>
[7] Disponível em <https://www.elsevier.com>
[8] Disponível em <https://www.ncbi.nlm.nih.gov/pubmed/>
[9] Disponível em <https://www.springer.com>
[10] Disponível em <https://library.maastrichtuniversity.nl/collections/databases/ssci/>
[11] Disponível em <https://link.springer.com>
[12] Disponível em <https://www.emerald.com/insight/>

Tabela 1 – Processo de extração

Filtro	Quantidade
Termos de busca	859
Filtros automáticos	−459
Leitura do abstract e palavras-chave	−71
Leitura completa	−300
TOTAL	30

Fonte: Autor (2019)

Tabela 2 – Critérios de avaliação da qualidade das publicações

ID	Critérios de avaliação	Respostas possíveis
C1	Independente de domínio	Sim = 1; Não = 0
C2	Seleção de atributos	Sim = 1; Não = 0
C3	Sensibilidade ao contexto	Sim = 1; Não = 0
C4	Avaliação empírica	Sim = 1; Não = 0
C5	Índice da revista	$\frac{FatordeImpacto2018}{MAX(FatordeImpacto2018)}$

Fonte: Autor (2019)

de fácil comparação das publicações selecionadas:

$$IndiceNormalizado = \frac{FatordeImpacto2018}{MAX(FatordeImpacto2018)} \tag{3.1}$$

A qualidade de cada ferramenta foi medida em percentual, que pode variar no intervalo de 0% a 100%, conforme definido na Equação 3.2. Este intervalo adota uma escala na qual valores próximos de 0% indicam que a ferramenta não atende aos critérios definidos nesta pesquisa, enquanto, os valores próximos de 100% indicam que a solução analisada atende aos critérios.

$$Qualidade = ((C1 + C2 + C3 + C4 + C5)/5) * 100 \tag{3.2}$$

Para selecionar as mais similares foi considerado que as ferramentas devem suportar pelo menos quatro critérios de qualidade, ou seja, atingir um percentual mínimo de 80% na avaliação de qualidade. Assim, dentre as 30 ferramentas levantadas durante a revisão da literatura, apenas 3 são relacionadas a esta pesquisa por atenderem aos critérios mínimos. Para mais detalhes sobre a avaliação dos trabalhos podem ser consultados no Apêndice .

3.2 Ferramentas mais similares

Nesta seção apresenta-se as seguintes ferramentas mais similares: *Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora*, *ADL* e

E2-P-D.

3.2.1 TR1: Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora

Este trabalho suporta a independência de domínio baseado na semântica formal dos dados descritos em RDF(S)/OWL, e também, estatísticas genéricas. Para realizar a seleção de atributos calcula a cardinalidade de cada atributo (quantidade de objetos distintos). A análise de contexto é baseado no grafo em que as triplas são inseridas no banco de triplas. Assim, a análise de contexto não será efetiva se a base de conhecimento de uma aplicação não estiver organizada em grafos, por isso, esta caraterística não pode ser considerada uma abordagem genérica para que consiga ser utilizada em qualquer aplicação. Além disso, realiza o alinhamento de atributos através da análise de correlação (HOGAN et al., 2012).

3.2.2 TR2: ADL

Este trabalho detecta inferência de relações de identidade usando um conjunto de propriedades importantes para efetuar a desambiguação de instâncias (referenciada como propriedades discriminativas) de forma automática. A aplicação desta abordagem é dividida em duas partes: *offline* e *online*. No primeiro, ele separa um conjunto de treinamento composto por dois grupos de pares de instâncias, mantendo seus links ou não, denominados de amostras positivas e negativas, respectivamente. A partir daí ele utiliza todas as triplas que possuem a URI de uma instância como sujeito para definir o *contexto* e determina a discriminalidade das propriedades através da ferramenta V-Doc (QU; HU; CHENG, 2006). Por fim, faz o emparelhamento de contextos e verifica quais propriedades possuem valores iguais nos contextos para determinar se há links entre as instâncias. Já a vinculação online recebe uma nova instância de entrada para extrair a classe e o nível de domínio. Em seguida, as classes de contrapartes e os níveis de domínios no conjunto de treinamento são escolhidos (HU; YANG; QU, 2014).

Pares de propriedades discriminativas são aprendidas através da comparação de contextos das instâncias no conjunto de treinamento, considerando que algumas propriedades são mais importantes que outras para caracterizar objetos do mundo real.

3.2.3 TR3: E2-P-D

Este trabalho apresenta um algoritmo para detectar entidades correferentes baseado em instâncias de classes comparáveis da ontologia. Para uma determinada instância, utiliza o processo de expansão encadeada para definir o contexto, encontrar o seu grafo de vizinhança e obter um conjunto de caminhos iniciando a partir desta instância e terminando em outro nó. Cada caminho é composto por várias triplas e para cada tripla é calculado a discriminabilidade baseado em seu predicado. Por fim, ele considera a distância (número de predicados) da tripla

até o nó raiz (a instância originária da busca pela vizinhança) associada a discriminabilidade para atribuir um peso a cada caminho do grafo de vizinhança (SONG; HEFLIN, 2013).

A discriminabilidade de um predicado representa os percentuais de variabilidade do sujeito e objeto, respectivamente. Para melhorar o desempenho de escalabilidade, após calcular a discriminabilidade para todos os predicados, ele aplica um filtro para considerar apenas os predicados que possuem discriminabilidade maior que um limiar pré-definido.

Por considerar que todas URIs distintas representam objetos diferentes no mundo real, ignorando URIs aliases, acaba superestimando o percentual de discriminabilidade, e consequentemente, aumenta a possibilidade de alcançar resultados imprecisos. Se o predicado *#hasAddress* alcança os objetos *#AddressA* e *#AddressB* – que são instâncias da entidade *Address* – a discriminabilidade dele vai ser 1. Mas, se assumirmos que estas duas instâncias de *Address* são equivalentes – representam o mesmo objeto – o algoritmo não é capaz de detectar este fato, por isso ocorre a superestimação mantendo a discriminabilidade em 1. Apesar de não ter avaliado o custo computacional para calcular a discriminabilidade das propriedades, esta pode representar uma outra limitação relevante, visto que a não utilização de um subconjunto de treinamento para calculá-la vai elevar o custo computacional porque é necessário recuperar todas as ocorrências do predicado a partir de todo o GC.

3.2.4 Síntese

Nesta seção é apresentada uma síntese das características dos trabalhos relacionados, conforme a Tabela 3. Não foi possível comparar o desempenho das soluções devido a indisponibilidade do código-fonte das soluções. Os *links* reportados nos artigos estão indisponíveis, tais como: <http://swse.deri.org/entity/> e <http://ws.nju.edu.cn/adl/>. Nas páginas pessoais dos autores na Web e nos seus grupos de pesquisa também não foi encontrado qualquer informação sobre o código fonte. Apesar disso, foram analisados os respectivos artigos científicos para extrair as características e comportamento, bem como, as vantagens e desvantagens de cada solução.

A seleção de atributos realizada por TR1, TR2 e TR3 é suscetível a superestimação da variabilidade, visto que ela sempre vai considerar URIs diferentes como recursos distintos pois não diferencia os objetos que já estiverem previamente identificados como similares. Além disso, nenhuma solução apresenta evidências sobre a redução de dimensionalidade no conjunto de dados. Uma comparação com a solução proposta será apresentada no Capítulo 5.

Tabela 3 – Síntese

Ferramenta	Seleção de atributos	Redução de dimensionalidade	Análise de contexto	Técnica de comparação
TR1	Objetos e sujeitos distintos	Não apresenta	Grafo (quadrupla)	Literal
TR2	Objetos distintos	Não apresenta	Caminho da instância	Literal
TR3	Objetos distintos	Não apresenta	Caminho da instância	String

Fonte: Autor (2019)

4 DESENVOLVIMENTO DA PESQUISA

Neste capítulo, discute-se sobre todas as etapas de desenvolvimento deste trabalho. O desenvolvimento foi baseado no processo proposto por (PEFFERS et al., 2007), executando as seguintes etapas: (i) *Identificação do problema*; (ii) *Objetivos*; (iii) *Projeto e Implementação*; e (iv) *Avaliação*.

4.1 Contexto da pesquisa

A dificuldade de inferir relações de identidade de forma automática foi reportada há nove anos por (HALPIN et al., 2010; ZAVERI et al., 2013). Recentemente, foi reforçada por (PAULHEIM, 2017) quando ele afirma que a geração de *links* entre dois GC é uma fonte de erro. Para compreender o problema, conhecer as soluções existentes e os desafios acerca deste tópico de pesquisa, foi realizada uma revisão da literatura (ver Capítulo 3) focada em IR.

O objetivo deste trabalho é melhorar o desempenho de inferência de relações de identidade, conforme a questão de pesquisa apresentada no Capítulo 1. Para isso, a referida questão de pesquisa pode ser dividida em duas, descritas a seguir:

> QP1: *A seleção automática de atributos melhora a acurácia de inferência de relações de identidade?*
>
> QP2: *A seleção automática de atributos reduz o espaço de processamento do motor de inferência?*

A melhora no desempenho pode ser alcançada com a evolução em aspectos *quantitativos* (avaliado na **QP1**) e *redução de dimensionalidade* (avaliado na **QP2**). A estratégia de inferência utilizada neste trabalho é uma inferência estatística focada na análise de variabilidade para seleção de atributos – que complementa a análise semântica – como um ponto chave para melhorar o desempenho durante a inferência de IR. Assim, a análise semântica está fora do escopo desta pesquisa.

4.2 Materiais e métodos

Os conjuntos de dados são provenientes de atividades de desenvolvimento de software relacionadas a 12 projetos reais que foram desenvolvidos pelo Laboratório de Sistemas Embarcados e Computação Pervasiva (Embedded) [1] da Universidade Federal de Campina Grande (UFCG)[2]. Eles foram compartilhados na estrutura de planilha e possuem 25 atributos que foram

[1] Disponível em <https://www.embedded.ufcg.edu.br>. Acessado em 13 de novembro de 2019.
[2] Disponível em <https://portal.ufcg.edu.br>. Acessado em 13 de novembro de 2019

coletados a partir das *sprints* executadas durante o desenvolvimento dos softwares, conforme Tabela 4.

Antes do compartilhamento para esta pesquisa, os dados passaram por um processo de anonimização[3] para que fossem removidas informações sigilosas que pudessem identificar os projetos (atributo *Projeto_ID*) e os colaboradores (atributo *Responsável*). Estas informações foram substituídas por identificadores genéricos, de tal modo que não comprometem o valor representado pelos atributos neste conjunto de dados por terem mantido o mesmo potencial de agregação do dado original.

Um especialista de domínio – que já atuou como Gerente de Projetos e Scrum Master durante 4 anos e como programador nas linguagens C#, Java e Android durante 2 anos – foi consultado para que ele pudesse apontar quais atributos são mais relevantes no conjunto de dados, atentando-se ao fato de que a relevância deve ser entendida como o potencial (força) que um atributo possui para determinar a identidade de um recurso existente no domínio e que o objetivo é utilizá-los para encontrar similaridade entre recursos. Ou seja, os atributos relevantes no conjunto de dados sobre tarefas de desenvolvimento de software devem ser capazes de distinguir unicamente as tarefas diferentes.

Dentre os 25 atributos, 14 (56%) foram definidos como relevantes pelo especialista do domínio (ver Tabela 4). Apesar de todos os projetos possuírem os mesmos atributos, eles possuem tamanhos (quantidade de tarefas) diferentes. Conforme apresentado na Tabela 5, a quantidade de tarefas varia de 45 a 212 por projeto.

4.2.1 Tecnologias utilizadas

Para implementação do código fonte[4] desta pesquisa foi utilizada a linguagem de programação Java 8 e as seguintes bibliotecas:

1. Apache POI[5]: auxilia no processamento das planilhas.
2. JDistlib[6]: para realização de testes de normalidade estatística.
3. Apache Commons Math 3[7]: para realizar a análise de correlação estatística.
4. Blazegraph[8]: banco de dados orientado a grafo.
5. Sesame API[9]: responsável pela comunicação com o Blazegraph.

[3] A anonimização é uma técnica de processamento de dados que remove ou modifica dados sigilosos.
[4] O códido fonte está disponível em: <https://github.com/williamsla/iri>
[5] Apache POI - the Java API for Microsoft Documents. Disponível em <https://poi.apache.org>
[6] JDistlib—Java Statistical Distribution Library. Disponível em <http://jdistlib.sourceforge.net>
[7] Commons Math: The Apache Commons Mathematics Library. Disponível em <https://commons.apache.org/proper/commons-math/>
[8] É um banco de dados orientado a grafo que fornece uma alta performance. Disponível em <https://blazegraph.com>
[9] <https://wiki.blazegraph.com/wiki/index.php/Sesame_API_Tutorial>

Tabela 4 – Descrição dos atributos

Atributo	Relevante	Descrição
Projeto_ID		Identificação do projeto
Plataforma	X	Ambiente de desenvolvimento utilizado
Arquitetura	X	Arquitetura utilizada no projeto
Domínio	X	Domínio de aplicação
Sprint_ID		Identificação da sprint
US_ID		Identificação da História do usuário
Tarefa_ID		Identificação da tarefa
Descritivo_da_US	X	Descrição da História do Usuário
Modulo	X	Componente que será alterado
Operacao	X	Ação executada pelo usuário no sistema
Tarefa_mapeada	X	Tarefa mapeada a partir da original
Tarefa_original	X	Tarefa original (texto livre)
Tags		Rótulos de categorização
Camada	X	Camada do software
Linguagem	X	Linguagem de programação
Framework	X	Framework utilizado para realizar a tarefa
API	X	API utilizada para realizar a tarefa
Persistencia	X	Banco de dados
Outras_Tags	X	Outros rótulos de categorização
Esforco_estimado_em_horas		Tempo estimado
Esforco_em_horas		Tempo gasto
NFR_Tipo		Tipo de requisito não-funcional
NFR_Atributo		Atributo do requisito não-funcional
NFR_Sentença		Sentença do requisito não-funcional
Responsavel		Responsável por realizar a tarefa

Fonte: Autor (2019)

4.3 Solução proposta: Um motor de inferência de relações de identidade

Nesta seção, apresenta-se a solução proposta neste trabalho. Como uma relação de identidade refere-se à percepção de que fatos identificados de formas diferentes podem referir-se a um mesmo recurso do mundo real, conforme discutido no Capítulo 1, é necessário contornar as dificuldades de detecção automática das restrições que podem determinar uma IR independente de domínio. Para isso, é necessário refletir sobre *como um atributo relevante pode ser identificado em um domínio*:

1. Qual o comportamento de um atributo relevante?
2. Há alguma relação entre atributos relevantes?

Fazendo uma analogia com distribuições de probabilidade, a análise do comportamento de um conjunto de dados deve atentar-se a duas perspectivas: i) as características predominantes

Tabela 5 – Quantidade de tarefas por projeto

Projeto	Quantidade de tarefas
P01	212
P02	112
P03	92
P04	88
P05	45
P06	61
P07	64
P08	53
P09	107
P10	48
P11	63
P12	117
TOTAL	1063

Fonte: Autor (2019)

na amostra; ii) as características excepcionais, que diferem-se da maioria. Com isso é possível compreender o comportamento e a relação entre as características. Normalmente é analisado se os atributos apresentam uma tendência crescente ou decrescente, se estão concentrados ou dispersos. No contexto de detecção automática de IR, a análise de comportamento refere-se também a capacidade (potencial) de um atributo ser um identificador em um domínio – similarmente a uma chave primária de um banco de dados, por exemplo. Assim, para desenvolver uma abordagem automática foram utilizadas técnicas de estatística e aprendizagem de máquina para auxiliar na seleção de atributos relevantes de um domínio a partir dos seus conjunto de valores.

A medição do potencial de identidade de um atributo ocorre a partir da análise de variabilidade e contexto a partir da análise de correlação. Enquanto a variabilidade mede a quantidade de informação representada por um atributo, o contexto analisa a relação entre atributos em um domínio. Com isso, evita-se a necessidade de que as restrições do domínio sejam determinadas previamente.

4.3.1 Projeto

Para realizar a inferência de IR a partir da seleção automática de atributos foram implementadas soluções para atuarem nos componentes *Motor de aprendizagem* e *Motor de inferência* – ilustrados na Figura 5.

4.3.1.1 Arquitetura

O componente *Seleção de atributos* é responsável por identificar os atributos mais relevantes do domínio de forma automática. Para isso ele é composto por três subcomponentes:

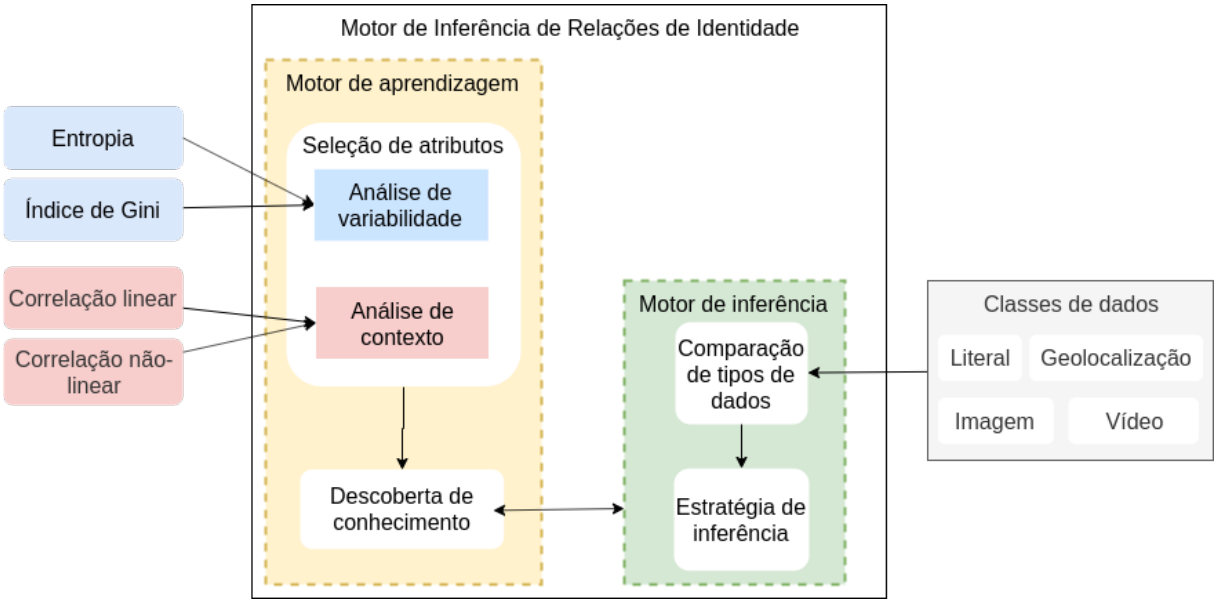

Figura 5 – Componente de seleção de atributos

Fonte: Autor (2019)

- *Análise de variabilidade*: indica o nível de relevância de um atributo – em uma escala de 0 a 1 – para inferir uma IR em um domínio. A relevância é avaliada através do cálculo da entropia dos dados, cujo objetivo é analisar a variação dos valores de um atributo.
- *Análise de contexto*: após a avaliação de variabilidade, realiza a análise de contexto através do estudo de correlação entre os atributos como um fator de ajuste a análise de variabilidade. Dessa forma, os atributos mais relevantes do domínio são selecionados.

O motor de inferência é responsável por analisar e comparar os registros do conjunto de dados para identificar similaridades. Ele é composto pelos seguintes subcomponentes:

- *Estratégia de comparação*: é uma generalização que suporta diversos algoritmos de comparação para as diversas classes de dados. Como cada classe possui suas particularidades, este componente é responsável por flexibilizar as formas de comparação dos valores de atributos através do encapsulamento de diferentes implementações. Por exemplo, para comparar textos podem ser acoplados neste componente as diversas abordagens disponíveis na literatura, tais como: Levenshtein Distance (YUJIAN; BO, 2007), Hamming Distance (NOROUZI; FLEET; SALAKHUTDINOV, 2012), e outros.
- *Estratégia de inferência*: abordagem genérica capaz de realizar inferências de relações de identidade a partir dos atributos relevantes selecionados.

4.3.1.2 Casos de uso

As funcionalidades da solução proposta são apresentadas com os casos de uso da Figura 6. Para alcançar o objetivo de realizar inferências de IR a partir da seleção de atributos, realiza-se as análises de variabilidade e de contexto. A variabilidade é obtida a partir da média entre a Entropia de Shannon e um cálculo simples de unicidade dos valores de um atributo – quantidade de valores únicos dividido pela quantidade total. O contexto é obtido a partir da análise de correlação, de tal modo que quando dois atributos possuírem alta correlação, então um deles é considerado irrelevante para determinar a identidade. Para isso, a correlação é calculada com o modelo de Pearson quando houver linearidade e o de Spearman no caso contrário.

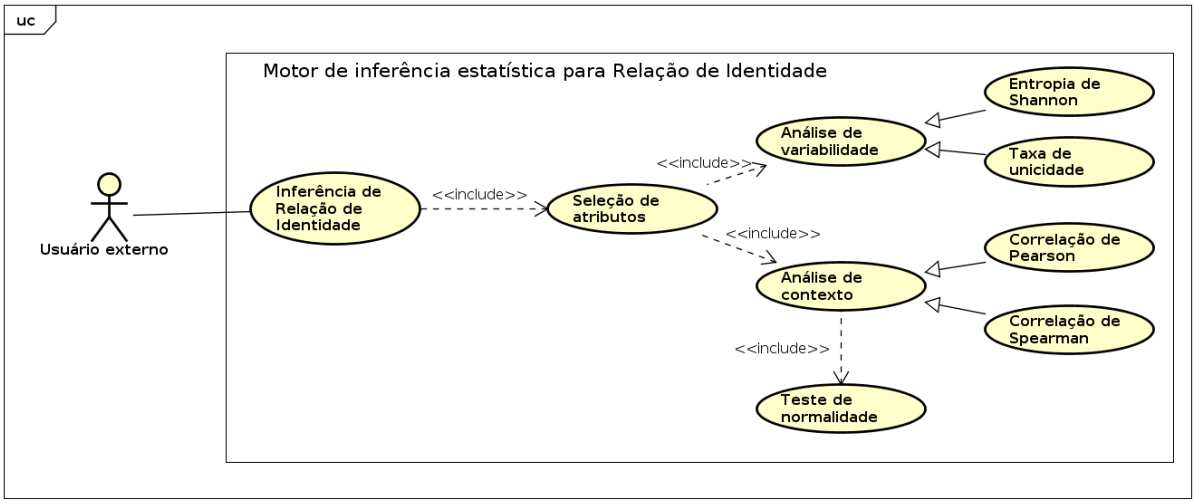

Figura 6 – Casos de uso

Fonte: Autor (2019)

4.3.2 Implementação

Foram utilizadas técnicas de aprendizagem de máquina e estatística já existentes na literatura para auxiliar na análise de distribuições de probabilidade associadas aos conjuntos de dados. Com a variedade de classes – texto, número, binário, data, geolocalização, etc – que os dados podem estar representados e a dificuldade de aplicar técnicas de estatísticas em classes de dados não-numéricas (nominais), foi necessário efetuar um mapeamento para uma representação equivalente na classe numérica decimal baseada na codificação ASCII (do inglês, American Standard Code for Information Interchange)[10], conforme o Trecho de código 4.1.

```
1     public Double textToNumeric(String value) {
2         Double result = 0d;
3
4         if (value != null) {
5
6             char[] chars = value.toCharArray();
7             for (char c : chars) {
```

[10] sua função é padronizar a forma como os computadores representam caracteres

```
8
9                   // convert from character to integer
10                  int convCharToInt = Character.getNumericValue(c);
11
12                  //ASCII does not supported
13                  if (convCharToInt < 0) {
14                      continue;
15                  }
16
17                  result += convCharToInt;
18              }
19          }
20
21          return result;
22      }
```

Listing 4.1 – Mapeando texto como número

4.3.2.1 Análise de variabilidade

O cálculo de relevância de um atributo é baseado na variabilidade dos valores dos atributos que é calculada a partir da Entropia de Shannon e a taxa de unicidade dos atributos.

A Entropia de Shannon, originada na teoria da informação, é capaz de medir a quantidade de incerteza associada a uma distribuição de probabilidade, ou seja, ela mede a quantidade de informação contida em uma variável. Inicialmente é realizado um agrupamento dos valores para seja contabilizado a ocorrência de repetições e aplicado o cálculo probabilístico de variabilidade de uma variável através da entropia, conforme o Trecho de código 4.2.

```
1  public class ShannonEntropy implements Entropy {
2
3      @Override
4      public <T> Double calculate(List<T> values) {
5          // count the occurrences of each value
6          Map<T, Long> mapClustering = values.stream().collect(
7                  Collectors.groupingBy(
8                          Function.identity(), Collectors.counting()
9                  )
10         );
11
12         Set<T> keys = mapClustering.keySet();
13
14         int k = (keys.size() <= 1) ? 2 : keys.size();
15
16         Double result = 0.0; //shannon entropy value
17         for (T key : keys) {
18             Double probability = (double) mapClustering.get(key) / values.
       size();
19             result -= probability * (Math.log(probability) / Math.log(k));
20         }
21
22         return result;
23     }
24 }
```

Listing 4.2 – Calculando a Entropia de Shannon

A entropia de Shannon tem sido comumente utilizada em variáveis que possuem valores binários, logo a quantidade de informação (*k*) é 2, seguindo a lógica do bit de informação discutida em (HEYLIGHEN; JOSLYN, 2001). Para exemplificar, vamos considerar que um conjunto de dados possui a variável *gênero* que pode conter dois valores: *masculino* ou *feminino*. Assim, qualquer registro do conjunto de dados só poderá ter um desses valores, portanto, a quantidade de informação contida nesta variável é 2 ($k = 2$). No entanto, como cada variável de um conjunto de dados pode representar coisas diferentes no mundo real, consequentemente, podem ter uma quantidade de informação diferente. Por isso, a abordagem automática deve ser capaz de identificar esta variação automaticamente, fazendo com que o cálculo de entropia adapte-se as particularidades de cada variável. Na linha 14 do Trecho de código 4.2, *k* é obtido a partir do agrupamento de informação da variável. Com isso, o valor da entropia de shannon para uma variável será entre 0 e 1, conforme definido na Equação 4.1, sendo *p* a frequência de ocorrência de cada valor da variável *X*.

$$H(X) = -\sum_{i=0}^{N-1} p_i log_k(p_i) \tag{4.1}$$

Apesar deste arcabouço já possuir a implementação do algoritmo para o cálculo da entropia de Shannon, este componente possui uma abstração polimórfica que permite o acoplamento de implementações externas, tais como o índice de Gini (DAGUM, 1980), para o cálculo da variabilidade que pode ser utilizada em substituição a entropia de Shannon.

A Taxa de Unicidade é utilizada como um fator de ajuste à Entropia de Shannon para determinar a relevância dos atributos. Ela é um índice mais geral que calcula a variação simples a partir da razão entre a quantidade de objetos distintos pela quantidade total de objetos de um atributo, conforme Equação 4.2.

$$F(X) = \frac{Quantidade\ de\ objetos\ distintos}{Quantidade\ total\ de\ objetos} \tag{4.2}$$

Assim, a relevância (R) dos atributos é determinada pela média aritmética entre a Entropia de Shannon e a Taxa de Unicidade, conforme Equação 4.3.

$$R(X) = \frac{H(X) + F(X)}{2} \tag{4.3}$$

4.3.2.2 Análise de contexto

Para complementar a avaliação da relevância, a análise de contexto é utilizada como a medida de correlação entre os atributos, a partir de dois algoritmos para análise de correlação: *Pearson* e *Spearman*.

- *Pearson*: aplicada em variáveis lineares de uma distribuição normal (gaussiana) de probabilidade.
- *Spearman*: é um teste não-paramétrico que mede o grau de associação entre duas variáveis.

Antes de aplicar a análise de correlação são realizados testes para identificar a distribuição de probabilidade associada a amostra para que seja possível aplicar a análise de correlação mais adequada. Inicialmente é aplicado o teste Shapiro-Wilk (SHAPHIRO; WILK, 1965) para verificar se as duas variáveis seguem uma distribuição normal. Se o resultado do teste indicar a normalidade, então será utilizada a correlação de Pearson. Caso contrário, a correlação de Spearman poderá ser utilizada, visto que ela independe da distribuição de probabilidade associada as amostras.

A análise de correlação poderá indicar a presença de contexto entre os atributos, de tal modo que uma atributo pertencerá ao contexto se ele estiver correlacionado com qualquer outro. Neste caso, o atributo que apresentar menor variabilidade, dentre os correlacionados, pode gerar redundância durante a análise de identidade e deverá ser considerado como irrelevante, visto que os atributos que são correlacionados com outros não introduzem qualquer novo conhecimento (JR et al., 2010).

4.3.2.3 Limiar para determinação de relevância

Como o cálculo da relevância (ver Equação 4.3) de atributos pode varia no intervalo de 0 a 1, é necessário determinar qual o valor mínimo (limiar) para considerar um atributo como relevante. Para isso, foi considerado o limiar que conseguisse maximizar as métricas de avaliação – *precision, recall, f-measure, accuracy, informedness e markedness* – ao realizar inferências de detecção de IR.

Conforme apresentado nas Figuras 7 e 8, todas as métricas atingem o melhor desempenho quando aplica-se o limiar igual a 0.5, alcançando rendimento satisfatório com valores próximos ao nível máximo. Isso significa que neste limiar o algoritmo de classificação consegue perceber mais informações positivas e negativas, e ainda, fornecer uma boa confiabilidade nas inferências.

Dessa forma, um atributo é considerado relevante quando o cálculo de variabilidade obedecer a Equação 4.4. Mais informações sobre a validação do limiar podem ser consultadas na Tabela 7 em Apêndice .

$$R(X) >= 0.5 \tag{4.4}$$

4.4 Avaliação

A avaliação da solução proposta foi baseada na eficácia das etapas: i) seleção automática de atributos; ii) inferências de IR. Para isso, foi realizado um estudo de caso, no qual, os dados

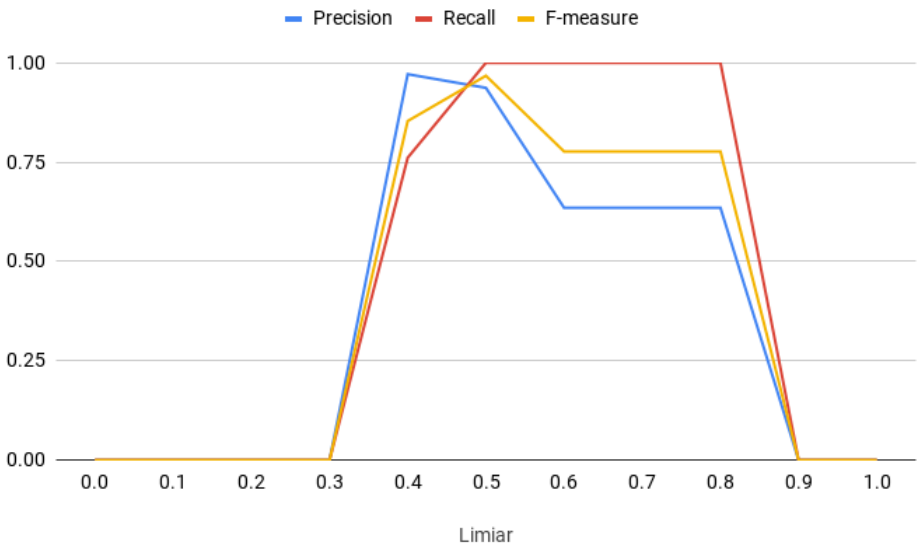

Figura 7 – Análise de limiar para predições positivas

Fonte: Autor (2019)

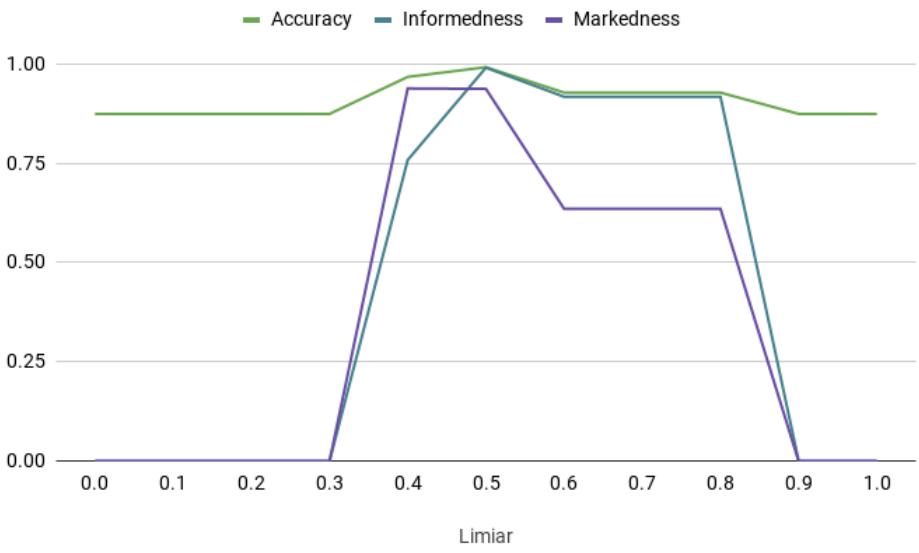

Figura 8 – Análise de limiar para predições positivas e negativas

Fonte: Autor (2019)

dos projetos foram submetidos separadamente e em conjunto ao motor de inferência estatística, gerando 13 cenários de avaliação – 12 projetos analisados individualmente e 1 análise conjunta de todos os projetos. Dessa forma possibilita a verificação da eficácia quando variar o tamanho da base de conhecimento, visto que em um pequeno conjunto de dados pode haver menos variabilidade e correlação entre os atributos. Além disso, em cada cenário haverá duas unidades de análise: i) seleção de atributos; ii) inferência de IR. A primeira pretende analisar a eficácia ao identificar atributos relevantes, enquanto a segunda refere-se a descoberta de recursos similares

no conjunto de dados. Logo, 26 unidades de análise serão avaliadas.

Como a análise semântica está fora do escopo deste trabalho, a conversão dos conjuntos de dados para um modelo de representação semântica, tal como OWL, poderia gerar um viés experimental já que as inferências de IR descobertas poderiam ser derivadas da análise semântica em vez da seleção de atributos relevantes realizada pelo motor de inferência. Assim, como o objetivo é avaliar se há melhora no desempenho de inferência de relações de identidade a partir da seleção automática de atributos, o estudo de caso foi executado diretamente a partir das planilhas.

Para avaliar o desempenho quantitativo em cada unidade de análise foram utilizadas as métricas de avaliação propostas por (POWERS, 2011): *precision*, *recall*, *F-measure*, *informedness* e *markedness*. O resultado obtido a partir destas métricas indicará o nível de confiabilidade das classificações realizadas pela solução proposta.

Para fins comparativos da acurácia na seleção automática de atributos, foi considerado que o conhecimento do especialista do domínio está correto, logo, desconsiderou-se eventuais influências causadas pela subjetividade da definição de relevância.

5 RESULTADOS E DISCUSSÃO

Este capítulo apresenta os resultados obtidos a partir da aplicação do estudo de caso em cada projeto, além de discuti-los e destacar algumas ameaças à validade da avaliação.

5.1 Resultados

Para facilitar o entendimento e sintetizar os resultados, nesta seção eles foram divididos em duas categorias: (i) seleção de atributos e (ii) inferência.

5.1.1 Seleção de atributos

Como a seleção de atributos utiliza análise estatística para caracterizar um atributo, a análise pode ser influenciada pelo tamanho do conjunto de dados. Conforme ilustrado na Tabela 6, na abordagem automática (colunas *P01, P02, P03, P04, P05, P06, P07, P08, P09, P10, P11, P12 e Todos*), a maioria dos atributos selecionados não são consenso entre os projetos. Baseando-se nos atributos relevantes definidos pelo especialista de domínio será discutido sobre a eficácia da solução em cada projeto.

Tabela 6 – Atributos selecionados por projeto

Atributos	Especialista	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	Todos
Projeto_ID														
Plataforma	X													
Arquitetura	X													
Domínio	X													
Sprint_ID								X		X		X		
US_ID		X	X	X	X	X	X	X	X	X	X	X	X	
Tarefa_ID					X	X			X		X	X	X	
Descritivo_da_US	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Modulo	X													
Operacao	X						X	X	X					
Tarefa_mapeada	X	X	X	X	X	X	X	X	X		X			
Tarefa_original	X	X	X	X	X	X	X	X	X	X	X	X		X
Camada	X													
Linguagem	X													
Framework	X													
API	X													
Persistencia	X													
Outras_Tags	X					X	X	X						
Esforco_estimado_em_horas											X			
Esforco_em_horas									X					
NFR_Tipo														
NFR_Atributo														
NFR_Sentença														
Responsavel								X	X					

Fonte: Autor (2019)

Dentre os 14 atributos relevantes definidos pelo especialista do domínio, apenas 5 foram identificados como relevantes em alguns cenários.

- *Descritivo_da_US*: P01, P02, P03, P04, P05, P06, P07, P08, P09, P10, P11, P12 e Todos
- *Operacao*: P06, P07 e P08
- *Tarefa_mapeada*: P01, P02, P03, P04, P05, P06, P07, P08 e P10
- *Tarefa_original*: P01, P02, P03, P04, P05, P06, P07, P08, P09, P10, P11 e Todos
- *Outas_Tags*: P05, P06 e P07

Em nenhum cenário os outros 9 atributos foram selecionados – *Plataforma, Arquitetura, Dominio, Modulo, Camada, Linguagem, Framework, API e Persistencia*. De acordo com os critérios estatísticos para análise de variabilidade, estes atributos não apresentam variação suficiente para capacitá-los como potenciais identificadores de uma tarefa quando avaliados de forma independente, pois as respectivas taxas de variabilidade são menores que 0.5 (ver Tabela 9 em Apêndice). Analisando os valores atribuídos nos conjuntos de dados, estes atributos possuem muitos valores repetidos, mesmo quando as tarefas são diferentes. Por exemplo, o atributo *Camada* pode receber os valores *Front-end ou Back-end*, então para cada tarefa diferente em um mesmo projeto os valores se repetirão com alta frequência. Da mesma forma ocorre para os demais atributos que não foram selecionados. Logo, isso é suficiente para determinar a não-relevância desse atributo para identificar uma IR no domínio.

Observando a frequência de seleção de atributos nos diferentes cenários de análise, é importante destacar quatro atributos que foram selecionados em pelo menos 9 $(69,23\%)$ cenários (ver Figura 9):

- *US_ID*: foi selecionado em 12 $(92,30\%)$ cenários, no entanto, todos possuem um pequeno conjunto de dados. Quando foi submetido ao cenário *Todos* que possui o maior conjunto de dados, ele não foi selecionado.
- *Descritivo_da_US*: foi selecionado em todos os cenários, indicando a sua alta relevância e que não sofreu influência do tamanho do conjunto de dados nos diferentes cenários.
- *Tarefa_mapeada*: foi selecionado em 9 $(69,23\%)$ cenários compostos por conjuntos de dados de tamanho pequeno.
- *Tarefa_original*: foi selecionado em 12 $(92,30\%)$ cenários, em apenas 1 $(7,69\%)$ pode ter sido influenciado pelo baixo tamanho da amostra desse conjunto de dados.

As métricas para analisar a eficácia foram calculadas a partir da contabilização de inferências positivas e negativas (discutidas no Capítulo 2).

Na avaliação ilustrada na Figura 10 para o conjunto de predições positivas, nos cenários cujo tamanho da base de dados é pequena, a métrica *precision* oscilou entre os níveis baixo e

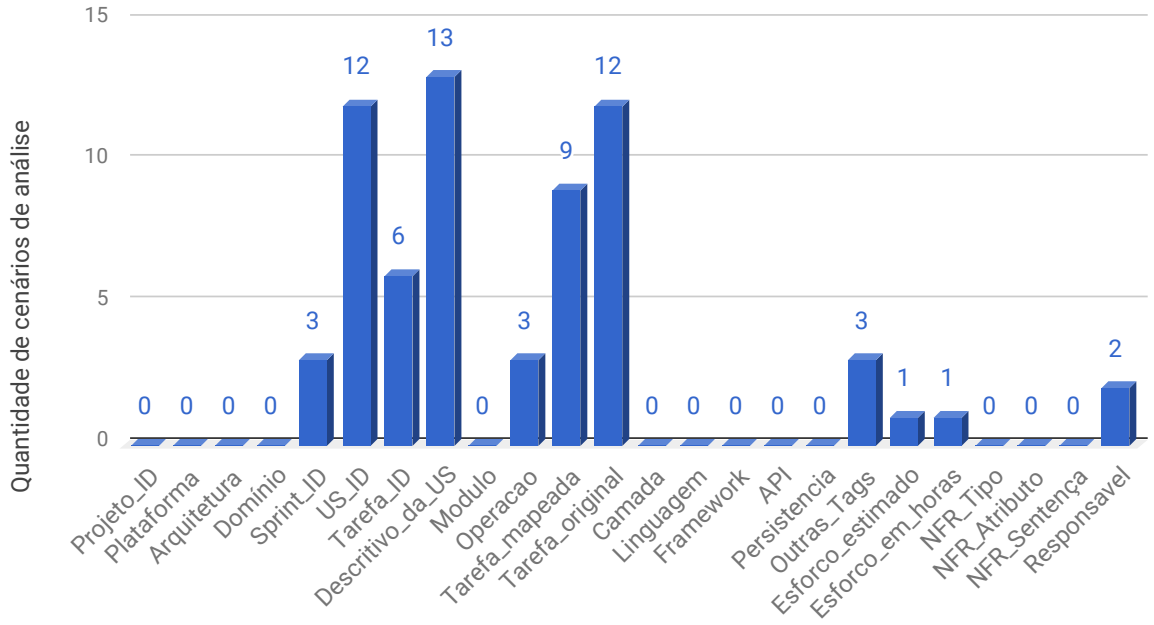

Figura 9 – Quantidade de cenários que os atributos foram selecionados

Fonte: Autor (2019)

intermediário – 0.33 a 0.83 – refletindo uma tendência de melhoria quando a solução é aplicada em conjuntos de dados maiores, conseguindo atingir o índice máximo (1.00). Isso indica que há um bom índice de acerto ao inferir que um atributo é relevante para um domínio. Já a *recall* manteve-se entre 0.11 e 0.40 que é considerada baixa, em todos os cenários, indicando que a solução possui baixa capacidade de percepção de atributos relevantes, visto que eles realmente existem na base. Ao comparar estas duas métricas através da *F-measure*, o índice mantém-se baixo, oscilando entre 0.17 e 0.5, além de apresentar uma tendência de decrescimento a medida que o tamanho da base aumenta. Logo, em termos gerais, a solução apresenta uma baixa capacidade de percepção de atributos relevantes, no entanto, possui alto índice de acertos ao realizar uma inferência de relevância positiva.

Para complementar com a avaliação das predições negativas, serão analisadas também as métricas *Accuracy, Informedness e Markedness*. A *Accuracy* mede a taxa total de acertos, enquanto a *Informedness* verifica a capacidade de percepção de informações positivas e negativas, e a *Markedness* refere-se a confiabilidade do classificador.

Na Figura 11, a *Accuracy* mantém-se constante em uma taxa intermediária próxima de 50% de acertos quando comparada com a seleção de atributos feita pelo especialista do domínio. A métrica *Informedness* mantém-se baixo e constante, independentemente do tamanho do conjunto de dados. Já a *Markedness* também mantém-se baixo para pequenas bases de dados, no entanto, apresenta uma tendência de crescimento à medida que o tamanho do conjunto de dados aumenta, apesar de ter alcançado um índice ainda considerado baixo (0.45) no cenário

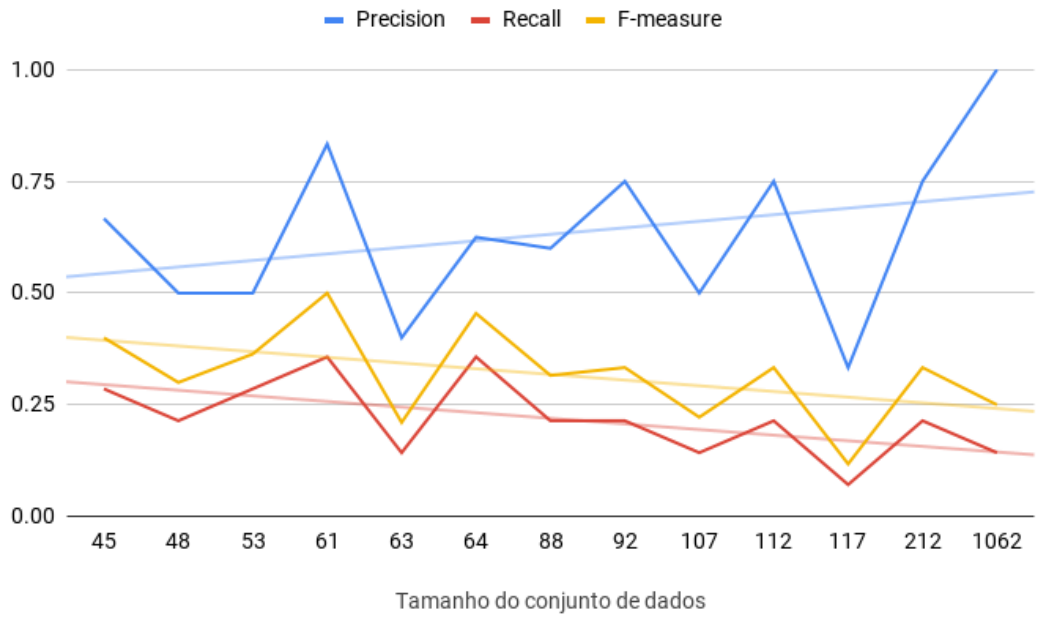

Figura 10 – Eficácia para predições positivas

Fonte: Autor (2019)

que possui mais quantidade de dados.

Por fim, dado o alcance do índice máximo de confiabilidade ao realizar inferências positivas de atributos relevantes no cenário com todos os conjuntos de dados, foi analisado a força de identidade para cada atributo. De acordo com a Figura 12, apenas os atributos *Tarefa_original* e *Descritivo_da_US* foram selecionados como relevantes para identificar uma IR no domínio, enquanto os demais apresentaram relevância abaixo do limiar (0.5). Com isso, é possível afirmar que estes atributos selecionados são realmente relevantes para o domínio, no entanto, é possível a existência de outros que também são relevantes.

5.1.2 Inferência estatística

A última etapa da avaliação da eficácia foi realizada a partir de predições de identidade utilizando duas abordagens: i) utilização de atributos selecionados pelo especialista; ii) seleção automática de atributos pelo IRI. Os dados das métricas de acurácia estão detalhados na Tabela 10 em Apêndice .

Na avaliação das predições positivas (presença de relação de identidade), o conjunto de atributos selecionados pelo IRI leva vantagem sob o do especialista, conforme indica a *F-measure* (0.97) representada na Figura 13. O especialista consegue melhor resultado na confiabilidade ao detectar uma relação de identidade, alcançando uma *precision* máxima de 1.00, enquanto o IRI alcança 0.94. No entanto, como o IRI apresenta maior índice de *recall*, ele possui maior capacidade de percepção de relações de identidade. Ou seja, o IRI consegue detectar mais

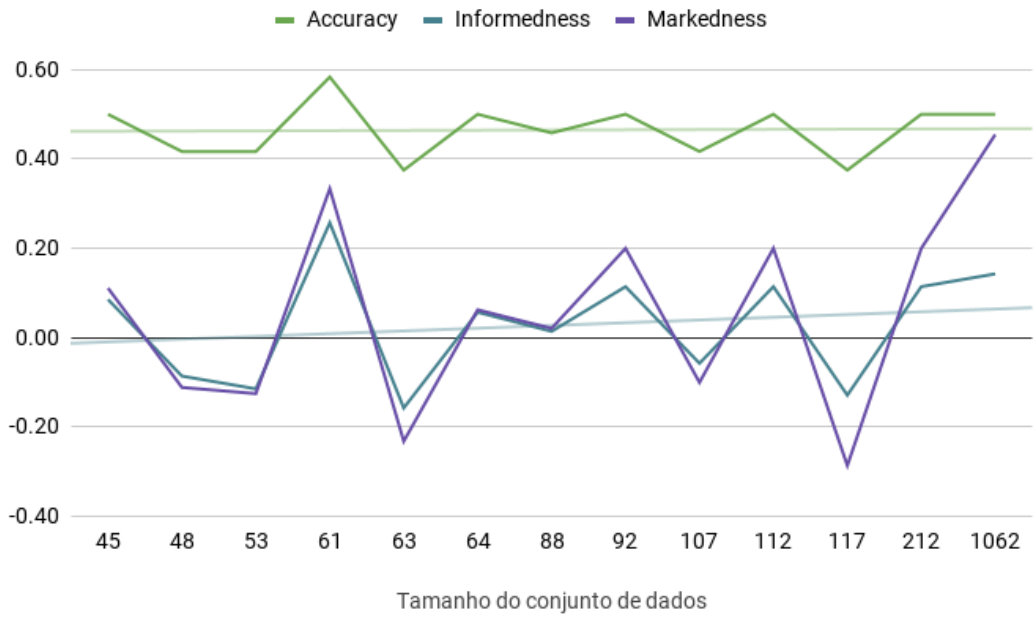

Figura 11 – Eficácia para predições positivas e negativas

Fonte: Autor (2019)

relações do que o Especialista, enquanto o Especialista consegue acertar mais que o IRI. No contexto geral (*F-measure*) de predições positivas, o IRI (0.97) leva uma pequena vantagem sob o Especialista (0.92).

Analisando em conjunto as predições positivas e negativas, de acordo com a Figura 14 na métrica *Informedness* o IRI (0.99) obteve melhor desempenho que o Especialista, indicando que a abordagem automática possui maior capacidade de percepção de informação. Enquanto na métrica *Markedness* o Especialista (0.98) foi superior ao IRI (0.94), apresentando maior confiabilidade ao realizar predições. De forma geral, o IRI (0.99) leva uma pequena vantagem sob o Especialista (0.98) na taxa geral de acertos (métrica *Accuracy*) ao realizar predições positivas e negativas.

5.2 Discussão

Os atributos que foram selecionados com alta frequência, em vários cenários, tem um alto potencial de relevância como identificador no domínio. Ficou evidente que quando a solução proposta é submetida a amostras pequenas, os índices de variabilidade estatística dos atributos podem ser influenciados, ocasionando a seleção de alguns atributos não-relevantes.

Apesar do IRI ter selecionado apenas 2 $(14,28\%)$ atributos relevantes dentre os que foram definidos pelo especialista do domínio, e consequentemente, este baixo percentual pode confundir-se com uma deficiência, o IRI obteve melhor acurácia ao realizar inferências de relações de identidade. Isso ocorre porque a determinação da relevância de atributos realizada

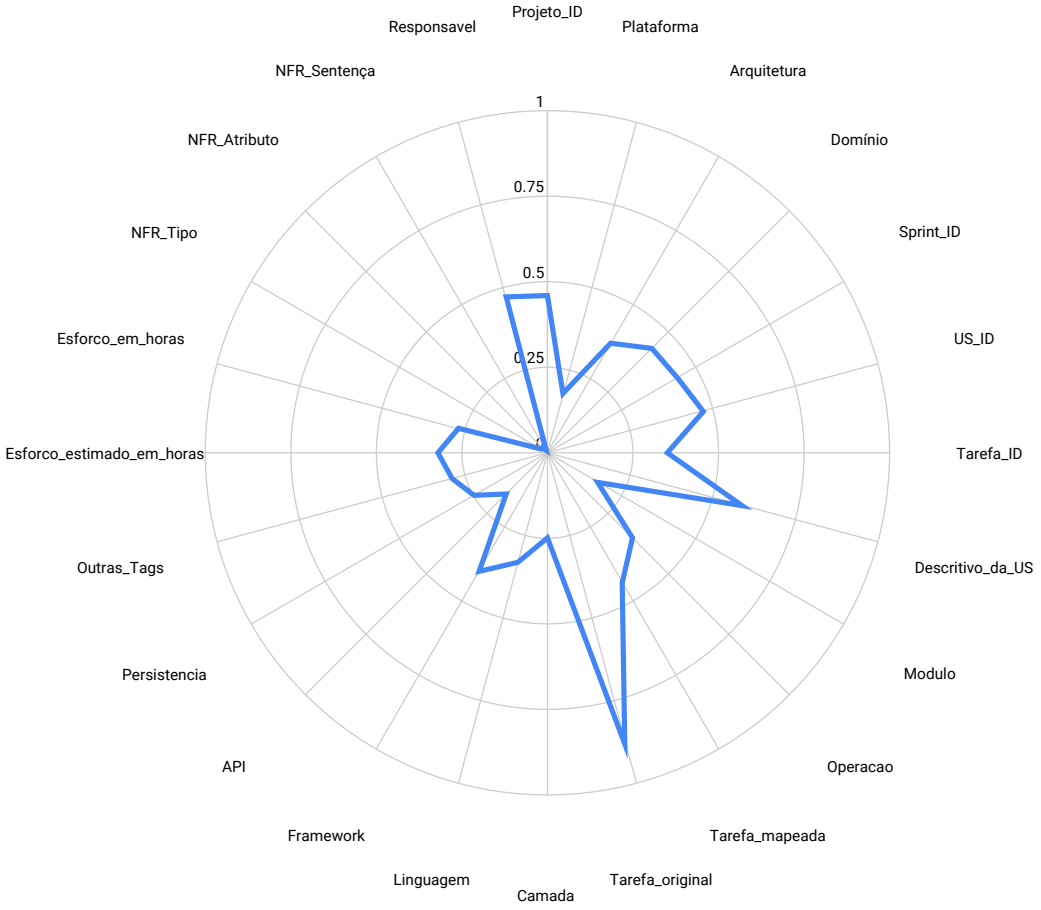

Figura 12 – Força dos atributos no cenário que possui todos os projetos

Fonte: Autor (2019)

por humanos pode ser influenciada pela subjetividade envolvida na tarefa de distinguir o atributo que é importante para complementar a informação sobre uma instância e o que é importante para identificá-la. No contexto de relação de identidade, um atributo só deve ser considerado relevante se ele tiver condições de identificar uma instância sem o complemento de outros atributos. Por exemplo, no domínio de *Pessoas*, os atributos *CPF* e *gênero* são relevantes, no entanto, apenas o *CPF* é suficiente para identificar uma instância deste domínio, independentemente de outros atributos dessa mesma instância. Se duas instâncias possuírem o mesmo valor para o CPF, então elas são iguais. Diferentemente do atributo *gênero* que, sozinho, não é capaz de identificar uma instância. Mesmo que duas instâncias possuam diferentes valores para o gênero, mas, se o CPF for igual, então elas serão consideradas iguais. Por isso, a seleção de atributos deve considerar que alguns atributos são mais relevantes que outros e analisar se há correlação entre eles. Dessa forma, deve-se avaliar a pertinência de determinar alguns atributos como relevantes no domínio de tarefas de desenvolvimento de software, tais como: *Plataforma, Arquitetura, Domínio, Camada, Framework, API, Persistencia e Outras_Tags*.

Como a avaliação é realizada a partir da análise do desempenho das duas abordagens (manual e automática) ao detectar IR, não se faz necessário apresentar o resultado da seleção de

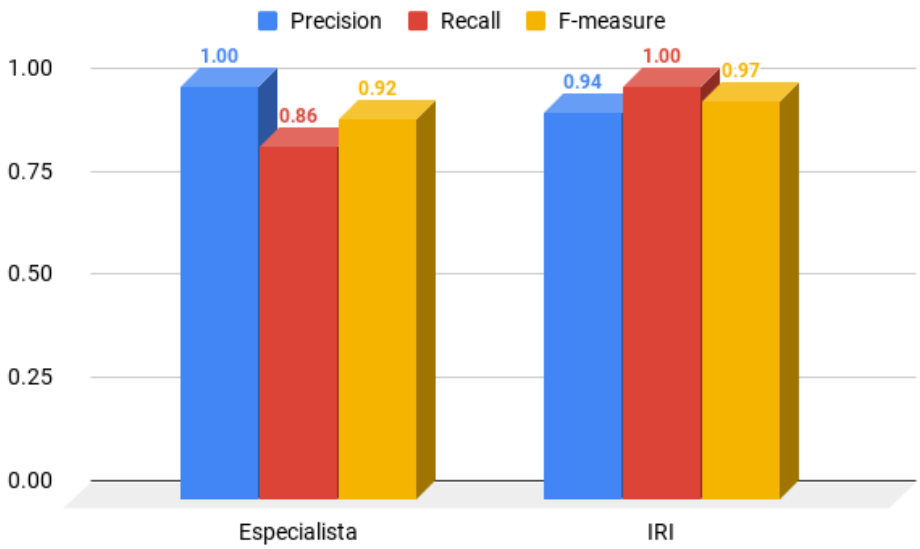

Figura 13 – Eficácia de para predições positivas entre Especialista e IRI

Fonte: Autor (2019)

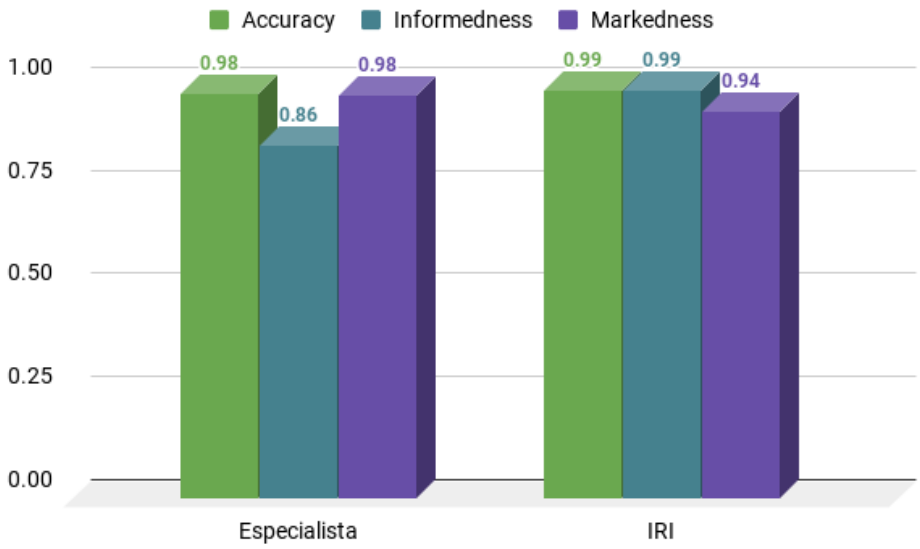

Figura 14 – Eficácia de predições positivas e negativas entre Especialista e IRI

Fonte: Autor (2019)

atributos feita pelo IRI para o especialista do domínio validá-las, visto que a análise das métricas de avaliação fornecem uma forma clara e objetiva de comparar e medir a validade das duas abordagens.

Apesar dos trabalhos relacionados (ver Capítulo 3) apresentarem várias características similares à solução proposta, não foi possível realizar uma comparação do desempenho das ferramentas quando aplicadas a um mesmo conjunto de dados porque não foi encontrado o código fonte de nenhum deles, conforme apresentado na Subseção 3.2.4 do Capítulo 3.

Diferentemente das estratégias existentes, esta solução baseia-se na Entropia de Shannon para obter o índice de variabilidade dos atributos. Com isso foi possível diminuir significativamente a dimensionalidade do conjunto de dados, e ainda, gerar melhores resultados na acurácia das inferências de relações de identidade. Além disso, é possível identificar que uma URI possui similares e, automaticamente, considerá-las como um mesmo recurso ao aplicar o cálculo da entropia de Shannon.

5.3 Ameaças à validade

Neste estudo de caso foram identificadas ameaças à validade de três tipos: construção, interna e externa. A ameaça de construção ocorre pela possibilidade da extração de dados realizada pelo colaborador que construiu o conjunto de dados não ter definido atributos suficientes para inferir relações de identidade. A ameaça interna ocorre pela possibilidade das relações de identidade não terem sido descobertas devido a influencia da seleção de atributos. Esta ameaça foi minimizada com a não utilização do banco de dados orientado a grafo para que não sofresse os efeitos dos raciocinadores semânticos.

A ameaça externa ocorre pelo fato do estudo de caso ter sido aplicado em um único domínio, por isso, não é possível afirmar que a solução proposta é válida em qualquer domínio. Outra ameaça externa é a subjetividade envolvida na seleção de atributos relevantes realizada pelo especialista de domínio, pois ele pode ter selecionado atributos que são irrelevantes ou pode não ter selecionado todos os relevantes. Para minimizar esta última ameaça foi explicado ao especialista que os atributos relevantes são considerados aqueles que possuem a capacidade de identificar um recurso no conjunto de dados.

6 CONCLUSÃO

Nesta dissertação, examinou-se o problema de baixa acurácia ao inserir *links* entre diferentes grafos de conhecimento, obtendo-se uma solução para realizar a seleção automática de atributos relevantes de um domínio através do uso de técnicas de análise de entropia e correlação estatística. A solução foi avaliada por um estudo de caso real, verificando-se a acurácia durante a inferência de IR com atributos selecionados de forma manual e automática. Assim, pode-se tirar a seguinte conclusão: com a aplicação das duas formas de seleção de atributos, percebeu-se que a solução proposta conseguiu efetuar mais inferências de IR com maior confiabilidade, apesar de ter selecionado menos atributos relevantes que o especialista do domínio. A acurácia inferior da abordagem manual pode ocorrer devido à subjetividade envolvida na análise do especialista acerca da definição de relevância de um atributo, visto que vários atributos podem ser necessários para definir uma IR em um domínio, dificultando a distinção entre a importância de complementar a descrição de um recurso e a de ser capaz de identificá-lo unicamente.

As duas abordagens possuem acurácias equivalentes durante a percepção da presença de IR, no entanto, para diminuir a ambiguidade é importante perceber também à ausência. Assim, analisando o contexto geral – com inferências positivas e negativas – a solução apresentada neste trabalho torna-se mais viável devido à superioridade da acurácia ao perceber as relações de identidade. A percepção de inferências negativas é uma característica importante para avaliar em qualquer software cujo objetivo é realizar inferências. Outro ponto importante foi a diminuição de dimensionalidade alcançada. A solução proposta reduziu 85.71% a quantidade de atributos selecionados e ainda conseguiu alcançar melhor acurácia durante as inferências.

6.1 Trabalhos futuros

O estudo de caso executado neste trabalho envolveu apenas um domínio de aplicação, impossibilitando a generalização quanto à validade da solução. Assim, evidentemente, é necessário aplicar em outros domínios para reunir mais evidências sobre a validade da independência de domínio.

O motor de inferência estatística desenvolvido nesta pesquisa considera que os dados foram construídos seguindo boas práticas de representação com utilização de valores padronizados, sem variação na escrita de dados similares, de tal modo que facilita para realização de uma extração de dados bem estruturada. No entanto, como no mundo real a construção de grandes conjuntos de dados é uma tarefa exaustiva para ser realizada de formal manual por humanos, é possível que a representação dos valores nos atributos sofra desvios na forma ou estrutura. Com isso, faz-se necessário a utilização de técnicas de processamento de linguagem natural para auxiliar na extração de dados para realizar as análises de similaridade.

Além disso, é relevante investigar acerca da combinação de atributos relevantes ao

determinar uma IR do domínio, baseando-se no conceito de chaves compostas (combinação de atributos) utilizado em bancos de dados relacionais para determinar que os respectivos atributos são capazes de identificar unicamente um recurso. Atualmente, os atributos são avaliados individualmente, sem levar em consideração a possibilidade de estimar a força em grupo para identificar uma IR.

Por fim, outro ponto importante refere-se a análise semântica. Ela deve ser adicionada a esta solução para atuar em conjunto com a análise sintática para detectar IR independente de domínio, podendo-se utilizar ontologias e motores de inferência dedutiva que operam em bancos de triplas.

REFERÊNCIAS

ACKOFF, R. L. From data to wisdom. *Journal of applied systems analysis*, v. 16, n. 1, p. 3–9, 1989. Citado 2 vezes nas páginas 9 e 10.

ALAM, M. et al. Event-based knowledge reconciliation using frame embeddings and frame similarity. *Knowledge-Based Systems*, Elsevier, v. 135, p. 192–203, 2017. Citado na página 54.

ARAUJO, S. et al. Serimi: Class-based disambiguation for effective instance matching over heterogeneous web data. In: *WebDB*. [S.l.: s.n.], 2012. p. 25–30. Citado na página 7.

AUER, S. et al. Dbpedia: A nucleus for a web of open data. In: *The semantic web*. [S.l.]: Springer, 2007. p. 722–735. Citado na página 16.

AUFAURE, M.-A. et al. From business intelligence to semantic data stream management. *Future Generation Computer Systems*, Elsevier, v. 63, p. 100–107, 2016. Citado na página 1.

BAADER, F.; HORROCKS, I.; SATTLER, U. Description logics. *Foundations of Artificial Intelligence*, Elsevier, v. 3, p. 135–179, 2008. Citado na página 6.

BARAL, C.; GELFOND, M. Logic programming and knowledge representation. *The Journal of Logic Programming*, Elsevier, v. 19, p. 73–148, 1994. Citado na página 10.

BEEK, W.; SCHLOBACH, S.; HARMELEN, F. van. A contextualised semantics for owl: sameas. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2016. p. 405–419. Citado na página 3.

BELLINGER, G.; CASTRO, D.; MILLS, A. Data, information, knowledge, and wisdom. 2004. Citado 2 vezes nas páginas 9 e 10.

BERNERS-LEE, T. Linked data-design issues. *http://www. w3. org/DesignIssues/LinkedData. html*, 2006. Citado 2 vezes nas páginas 12 e 13.

BERNERS-LEE, T.; FIELDING, R.; MASINTER, L. *Uniform resource identifier (URI): Generic syntax*. [S.l.], 2004. Citado na página 13.

BERNERS-LEE, T. et al. The semantic web. *Scientific american*, New York, NY, USA:, v. 284, n. 5, p. 28–37, 2001. Citado 2 vezes nas páginas 1 e 11.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked data: The story so far. In: *Semantic services, interoperability and web applications: emerging concepts*. [S.l.]: IGI Global, 2011. p. 205–227. Citado 5 vezes nas páginas 1, 2, 12, 13 e 16.

BLUMAUER, A. *From taxonomies over ontologies to knowledge graphs*. 2014. Citado 2 vezes nas páginas 1 e 2.

BORDES, A. et al. A semantic matching energy function for learning with multi-relational data. *Machine Learning*, Springer, v. 94, n. 2, p. 233–259, 2014. Citado na página 14.

BOTTA, A. et al. Integration of cloud computing and internet of things: a survey. *Future generation computer systems*, Elsevier, v. 56, p. 684–700, 2016. Citado na página 1.

BOVET, A.; MAKSE, H. A. Influence of fake news in twitter during the 2016 us presidential election. *Nature communications*, Nature Publishing Group, v. 10, n. 1, p. 7, 2019. Citado na página 9.

BU, Z.; XIA, Z.; WANG, J. A sock puppet detection algorithm on virtual spaces. *Knowledge-Based Systems*, Elsevier, v. 37, p. 366–377, 2013. Citado na página 9.

CAI, H.; ZHENG, V. W.; CHANG, K. C.-C. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 30, n. 9, p. 1616–1637, 2018. Citado na página 4.

CHEIN, M.; MUGNIER, M.-L. *Graph-based knowledge representation: computational foundations of conceptual graphs*. [S.l.]: Springer Science & Business Media, 2008. Citado na página 14.

CHEN, C. P.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information sciences*, Elsevier, v. 275, p. 314–347, 2014. Citado na página 1.

CHEN, L. et al. Ahab: Aligning heterogeneous knowledge bases via iterative blocking. *Information Processing & Management*, Elsevier, v. 56, n. 1, p. 1–13, 2019. Citado na página 54.

CHENG, G.; XU, D.; QU, Y. C3d+p: A summarization method for interactive entity resolution. *Journal of Web Semantics*, Elsevier BV, v. 35, p. 203–213, dez. 2015. Disponível em: <https://doi.org/10.1016/j.websem.2015.05.004>. Citado na página 54.

COLUCCI, S. et al. Defining and computing least common subsumers in rdf. *Journal of Web Semantics*, Elsevier, v. 39, p. 62–80, 2016. Citado na página 54.

DAGUM, C. Inequality measures between income distributions with applications. *Econometrica (pre-1986)*, Blackwell Publishing Ltd., v. 48, n. 7, p. 1791, 1980. Citado na página 31.

DAVIS, R.; SHROBE, H.; SZOLOVITS, P. What is a knowledge representation? *AI magazine*, v. 14, n. 1, p. 17–17, 1993. Citado 2 vezes nas páginas 3 e 10.

DEMARTINI, G.; DIFALLAH, D. E.; CUDRÉ-MAUROUX, P. Large-scale linked data integration using probabilistic reasoning and crowdsourcing. *The VLDB Journal*, v. 22, p. 665–687, 2013. Citado na página 54.

DEY, A. K. Understanding and using context. *Personal and ubiquitous computing*, Springer-Verlag, v. 5, n. 1, p. 4–7, 2001. Citado na página 13.

DONG, X. L.; SRIVASTAVA, D. Big data integration. In: IEEE. *2013 IEEE 29th international conference on data engineering (ICDE)*. [S.l.], 2013. p. 1245–1248. Citado na página 1.

DYLLA, F. et al. A survey of qualitative spatial and temporal calculi: algebraic and computational properties. *ACM Computing Surveys (CSUR)*, ACM, v. 50, n. 1, p. 7, 2017. Citado na página 3.

EHRLINGER, L.; WÖSS, W. Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, v. 48, 2016. Citado na página 1.

FÄRBER, M. et al. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web*, IOS Press, v. 9, n. 1, p. 77–129, 2018. Citado 2 vezes nas páginas 1 e 2.

FIELDING, R. et al. *Hypertext transfer protocol–HTTP/1.1*. [S.l.], 1999. Citado na página 13.

GALLINUCCI, E. et al. Interactive multidimensional modeling of linked data for exploratory olap. *Information Systems*, Elsevier, v. 77, p. 86–104, 2018. Citado na página 54.

GANGEMI, A. et al. Identifying motifs for evaluating open knowledge extraction on the web. *Knowledge-Based Systems*, Elsevier, v. 108, p. 33–41, 2016. Citado na página 54.

GOYAL, P.; FERRARA, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems*, Elsevier, v. 151, p. 78–94, 2018. Citado na página 4.

GRAHAM, R. Google and advertising: digital capitalism in the context of post-fordism, the reification of language, and the rise of fake news. *Palgrave Communications*, Nature Publishing Group, v. 3, n. 1, p. 45, 2017. Citado na página 9.

GROSAN, C.; ABRAHAM, A. Knowledge representation and reasoning. In: *Intelligent Systems*. [S.l.]: Springer, 2011. p. 131–147. Citado na página 10.

GROUP, W. S. working et al. Sparql 1.1 overview. *W3C Recommendation. W3C*, 2013. Citado na página 13.

GROZA, T. et al. From raw publications to linked data. *Knowledge and Information Systems*, v. 34, p. 1–21, 2011. Citado na página 54.

GUAN, N.; SONG, D.; LIAO, L. Knowledge graph embedding with concepts. *Knowledge-Based Systems*, Elsevier, v. 164, p. 38–44, 2019. Citado 2 vezes nas páginas 1 e 2.

GUARINO, N. Formal ontology, conceptual analysis and knowledge representation. *International journal of human-computer studies*, Elsevier, v. 43, n. 5-6, p. 625–640, 1995. Citado na página 10.

HALPIN, H. et al. When owl:sameas isn't the same: An analysis of identity in linked data. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2010. p. 305–320. Citado 2 vezes nas páginas 3 e 24.

HALPIN, H.; MCNEILL, F. Discovering meaning on the go in large heterogenous data. *Artificial Intelligence Review*, Springer, v. 40, n. 2, p. 107–126, 2013. Citado 2 vezes nas páginas 3 e 7.

HASSANZADEH, O.; CONSENS, M. P. Linked movie data base. In: *LDOW*. [S.l.: s.n.], 2009. Citado na página 16.

HEATH, T. Linked data-welcome to the data network. *IEEE Internet Computing*, IEEE, v. 15, n. 6, p. 70–73, 2011. Citado na página 12.

HEATH, T.; BIZER, C. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, Morgan & Claypool Publishers, v. 1, n. 1, p. 1–136, 2011. Citado na página 1.

HEYLIGHEN, F.; JOSLYN, C. Cybernetics and second-order cybernetics. *Encyclopedia of physical science & technology*, v. 4, p. 155–170, 2001. Citado na página 31.

HOEKSTRA, R. Ontology representation design patterns and ontologies that make sense. In: IOS PRESS. *Proceedings of the 2009 conference on Ontology Representation: Design Patterns and Ontologies that Make Sense*. [S.l.], 2009. p. 1–236. Citado 2 vezes nas páginas 10 e 11.

HOGAN, A. et al. Scalable and distributed methods for entity matching, consolidation and disambiguation over linked data corpora. *J. Web Semant.*, v. 10, p. 76–110, 2012. Citado 2 vezes nas páginas 22 e 54.

HORROCKS, I.; SATTLER, U.; TOBIES, S. Practical reasoning for expressive description logics. In: SPRINGER. *International Conference on Logic for Programming Artificial Intelligence and Reasoning*. [S.l.], 1999. p. 161–180. Citado na página 6.

HU, W.; JIA, C. A bootstrapping approach to entity linkage on the semantic web. *Journal of Web Semantics*, Elsevier BV, v. 34, p. 1–12, out. 2015. Disponível em: <https://doi.org/10.1016/j.websem.2015.07.003>. Citado 2 vezes nas páginas 3 e 54.

HU, W.; YANG, R.; QU, Y. Automatically generating data linkages using class-based discriminative properties. *Data & Knowledge Engineering*, v. 91, p. 34–51, 2014. Citado 2 vezes nas páginas 22 e 54.

JOHNSON-LAIRD, P. N. Deductive reasoning. *Annual review of psychology*, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 50, n. 1, p. 109–135, 1999. Citado na página 6.

JR, C. T. et al. Fast feature selection using fractal dimension. *Journal of Information and data Management*, v. 1, n. 1, p. 3–3, 2010. Citado na página 32.

KEJRIWAL, M.; MIRANKER, D. P. An unsupervised instance matcher for schema-free RDF data. *Journal of Web Semantics*, Elsevier BV, v. 35, p. 102–123, dez. 2015. Disponível em: <https://doi.org/10.1016/j.websem.2015.07.002>. Citado na página 54.

KLYNE, G.; CARROLL, J. J. Resource description framework (rdf): Concepts and abstract syntax. 2006. Citado na página 13.

LEAKE, D.; MAGUITMAN, A.; REICHHERZER, T. Experience-based support for human-centered knowledge modeling. *Knowledge-based systems*, Elsevier, v. 68, p. 77–87, 2014. Citado na página 10.

LI, J. et al. Rimom: A dynamic multistrategy ontology alignment framework. *IEEE Transactions on Knowledge and data Engineering*, IEEE, v. 21, n. 8, p. 1218–1232, 2009. Citado na página 7.

LIN, Y. et al. Learning entity and relation embeddings for knowledge graph completion. In: *Twenty-ninth AAAI conference on artificial intelligence*. [S.l.: s.n.], 2015. Citado 2 vezes nas páginas 1 e 2.

LIU, W. et al. A new truth discovery method for resolving object conflicts over linked data with scale-free property. *Knowledge and Information Systems*, Springer, v. 59, n. 2, p. 465–495, 2019. Citado na página 7.

LIU, W. et al. Representation learning over multiple knowledge graphs for knowledge graphs alignment. *Neurocomputing*, Elsevier, v. 320, p. 12–24, 2018. Citado na página 54.

MEYMANDPOUR, R.; DAVIS, J. G. A semantic similarity measure for linked data: An information content-based approach. *Knowledge-Based Systems*, Elsevier, v. 109, p. 276–293, 2016. Citado na página 54.

MONGIOVÌ, M. et al. Merging open knowledge extracted from text with mergilo. *Knowledge-Based Systems*, Elsevier, v. 108, p. 155–167, 2016. Citado na página 54.

MOUNTANTONAKIS, M.; TZITZIKAS, Y. Scalable methods for measuring the connectivity and quality of large numbers of linked datasets. *Journal of Data and Information Quality (JDIQ)*, ACM, v. 9, n. 3, p. 15, 2018. Citado na página 54.

NARDUCCI, F.; PALMONARI, M.; SEMERARO, G. Cross-lingual link discovery with tr-esa. *Information Sciences*, Elsevier, v. 394, p. 68–87, 2017. Citado na página 54.

NGOMO, A.-C. N. On link discovery using a hybrid approach. *Journal on Data Semantics*, v. 1, p. 203–217, 2012. Citado na página 54.

NOROUZI, M.; FLEET, D. J.; SALAKHUTDINOV, R. R. Hamming distance metric learning. In: *Advances in neural information processing systems*. [S.l.: s.n.], 2012. p. 1061–1069. Citado na página 28.

OLIVEIRA, B. C. et al. Ontogenesis: an architecture for automatic semantic enhancement of data services. *International Journal of Web Information Systems*, Emerald Publishing Limited, v. 15, n. 1, p. 2–27, 2019. Citado na página 54.

PARIS, P.-H. Assessing the quality of owl: sameas links. In: SPRINGER. *European Semantic Web Conference*. [S.l.], 2018. p. 304–313. Citado na página 6.

PAULHEIM, H. Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, IOS Press, v. 8, n. 3, p. 489–508, 2017. Citado 6 vezes nas páginas 1, 2, 3, 4, 16 e 24.

PEFFERS, K. et al. A design science research methodology for information systems research. *Journal of management information systems*, Taylor & Francis, v. 24, n. 3, p. 45–77, 2007. Citado na página 24.

PERNELLE, N.; SAÏS, F.; SYMEONIDOU, D. An automatic key discovery approach for data linking. *J. Web Semant.*, v. 23, p. 16–30, 2013. Citado na página 54.

PFAFF, M.; KRCMAR, H. A web-based system architecture for ontology-based data integration in the domain of it benchmarking. *Enterprise Information Systems*, Taylor & Francis, v. 12, n. 3, p. 236–258, 2018. Citado na página 54.

PILEGGI, S. F. Web of similarity. *Journal of Computational Science*, Elsevier, 2016. Citado na página 54.

POOLE, D. L.; MACKWORTH, A. K. *Artificial Intelligence: foundations of computational agents*. [S.l.]: Cambridge University Press, 2010. Citado na página 14.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. Bioinfo Publications, 2011. Citado 2 vezes nas páginas 17 e 34.

PUJARA, J. et al. Knowledge graph identification. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2013. p. 542–557. Citado 2 vezes nas páginas 1 e 2.

QU, Y.; HU, W.; CHENG, G. Constructing virtual documents for ontology matching. In: ACM. *Proceedings of the 15th international conference on World Wide Web*. [S.l.], 2006. p. 23–31. Citado na página 22.

RAAD, J. et al. Detecting erroneous identity links on the web using network metrics. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2018. p. 391–407. Citado na página 3.

RAAD, J.; PERNELLE, N.; SAÏS, F. Detection of contextual identity links in a knowledge base. In: ACM. *Proceedings of the Knowledge Capture Conference*. [S.l.], 2017. p. 8. Citado na página 3.

REBELE, T. et al. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2016. p. 177–185. Citado na página 16.

REED, S. K.; PEASE, A. Reasoning from imperfect knowledge. *Cognitive Systems Research*, Elsevier, v. 41, p. 56–72, 2017. Citado 2 vezes nas páginas 3 e 6.

RISTOSKI, P.; BIZER, C.; PAULHEIM, H. Mining the web of linked data with RapidMiner. *Journal of Web Semantics*, Elsevier BV, v. 35, p. 142–151, dez. 2015. Disponível em: <https://doi.org/10.1016/j.websem.2015.06.004>. Citado na página 54.

RIZZO, G.; ERP, M. van; TRONCY, R. Benchmarking the extraction and disambiguation of named entities on the semantic web. In: *LREC*. [S.l.: s.n.], 2014. p. 4593–4600. Citado na página 7.

SALVADORI, I. L. et al. Improving entity linking with ontology alignment for semantic microservices composition. *International Journal of Web Information Systems*, Emerald Publishing Limited, v. 13, n. 3, p. 302–323, 2017. Citado na página 54.

SCHILIT, B. N. et al. *Context-aware computing applications*. [S.l.]: Xerox Corporation, Palo Alto Research Center, 1994. Citado 2 vezes nas páginas 13 e 14.

SCHMACHTENBERG, M.; BIZER, C.; PAULHEIM, H. Adoption of the linked data best practices in different topical domains. In: SPRINGER. *International Semantic Web Conference*. [S.l.], 2014. p. 245–260. Citado na página 16.

ŠESTÁK, J.; MAREŠ, J. J.; HUBÍK, P. *Glassy, amorphous and nano-crystalline materials: thermal physics, analysis, structure and properties*. [S.l.]: Springer Science & Business Media, 2010. v. 8. Citado na página 17.

SHANNON, C. E. A mathematical theory of communication. *Bell system technical journal*, Wiley Online Library, v. 27, n. 3, p. 379–423, 1948. Citado na página 16.

SHAO, C. et al. Rimom-im: A novel iterative framework for instance matching. *Journal of computer science and technology*, Springer, v. 31, n. 1, p. 185–197, 2016. Citado na página 54.

SHAPHIRO, S.; WILK, M. An analysis of variance test for normality. *Biometrika*, v. 52, n. 3, p. 591–611, 1965. Citado na página 32.

SLEEMAN, J.; FININ, T.; JOSHI, A. Entity type recognition for heterogeneous semantic graphs. *AI Magazine*, Association for the Advancement of Artificial Intelligence (AAAI), v. 36, n. 1, p. 75, mar. 2015. Disponível em: <https://doi.org/10.1609/aimag.v36i1.2569>. Citado na página 54.

SONG, D.; HEFLIN, J. Domain-independent entity coreference for linking ontology instances. *Journal of Data and Information Quality (JDIQ)*, v. 4, p. 7:1–7:29, 2013. Citado 2 vezes nas páginas 23 e 54.

SONG, F.; ZACHAREWICZ, G.; CHEN, D. An ontology-driven framework towards building enterprise semantic information layer. *Advanced Engineering Informatics*, v. 27, p. 38–50, 2013. Citado na página 54.

STERCKX, L. et al. Knowledge base population using semantic label propagation. *Knowledge-Based Systems*, Elsevier, v. 108, p. 79–91, 2016. Citado na página 10.

TRONCY, R.; MALOCHA, B.; FIALHO, A. T. Linking events with media. In: ACM. *Proceedings of the 6th international conference on semantic systems*. [S.l.], 2010. p. 42. Citado na página 16.

WANG, Q. et al. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 29, n. 12, p. 2724–2743, 2017. Citado 2 vezes nas páginas 3 e 4.

WANG, S.; ENGLEBIENNE, G.; SCHLOBACH, S. Learning concept mappings from instance similarity. In: SPRINGER. *International semantic web conference*. [S.l.], 2008. p. 339–355. Citado na página 3.

WANG, Y.; KUNG, L.; BYRD, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, Elsevier, v. 126, p. 3–13, 2018. Citado na página 1.

YUJIAN, L.; BO, L. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 29, n. 6, p. 1091–1095, 2007. Citado na página 28.

ZAVERI, A. et al. User-driven quality evaluation of dbpedia. In: ACM. *Proceedings of the 9th International Conference on Semantic Systems*. [S.l.], 2013. p. 97–104. Citado 2 vezes nas páginas 3 e 24.

ZHENG, J. G. et al. Sem+: tool for discovering concept mapping in earth science related domain. *Earth Science Informatics*, v. 8, n. 1, p. 95–102, Mar 2015. ISSN 1865-0481. Disponível em: <https://doi.org/10.1007/s12145-014-0203-1>. Citado na página 54.

ZONG, N. et al. Aligning ontologies with subsumption and equivalence relations in linked data. *Knowledge-Based Systems*, Elsevier BV, v. 76, p. 30–41, mar. 2015. Disponível em: <https://doi.org/10.1016/j.knosys.2014.11.022>. Citado na página 54.

APÊNDICE A: MÉTRICAS DE AVALIAÇÃO

Tabela 7 – Validação do limiar para determinar a relevância de um atributo

Limiar	VP	FN	FP	VN	Precision	Recall	F-measure	Accuracy	Informedness	Markedness
0.0	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00
0.1	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00
0.2	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00
0.3	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00
0.4	102	32	3	926	0.97	0.76	0.85	0.97	0.76	0.94
0.5	134	0	9	920	0.94	1.00	0.97	0.99	0.99	0.94
0.6	134	0	77	852	0.64	1.00	0.78	0.93	0.92	0.64
0.7	134	0	77	852	0.64	1.00	0.78	0.93	0.92	0.64
0.8	134	0	77	852	0.64	1.00	0.78	0.93	0.92	0.64
0.9	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00
1.0	0	134	0	929	0.00	0.00	0.00	0.87	0.00	0.00

Tabela 8 – Parâmetros para cálculo das métricas de avaliação da seleção de atributos

Cenário	Quantidade de tarefas	VP	FN	FP	VN	Precision	Recall	F-measure	Accuracy	Informedness	Markedness
P01	212	3	11	1	9	0.75	0.21	0.33	0.50	0.11	0.20
P02	112	3	11	1	9	0.75	0.21	0.33	0.50	0.11	0.20
P03	92	3	11	1	9	0.75	0.21	0.33	0.50	0.11	0.20
P04	88	3	11	2	8	0.60	0.21	0.32	0.46	0.01	0.02
P05	45	4	10	2	8	0.67	0.29	0.40	0.50	0.09	0.11
P06	61	5	9	1	9	0.83	0.36	0.50	0.58	0.26	0.33
P07	64	5	9	3	7	0.63	0.36	0.45	0.50	0.06	0.06
P08	53	4	10	4	6	0.50	0.29	0.36	0.42	-0.11	-0.13
P09	107	2	12	2	8	0.50	0.14	0.22	0.42	-0.06	-0.10
P10	48	3	11	3	7	0.50	0.21	0.30	0.42	-0.09	-0.11
P11	63	2	12	3	7	0.40	0.14	0.21	0.38	-0.16	-0.23
P12	117	1	13	2	8	0.33	0.07	0.12	0.38	-0.13	-0.29
Todos	1062	2	12	0	10	1.00	0.14	0.25	0.50	0.14	0.45

Tabela 9 – Taxa de variabilidade

Atributos	Especialista	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10	P11	P12	Todos
Projeto_ID	0.00	0.03	0.06	0.07	0.07	0.13	0.10	0.10	0.11	0.06	0.12	0.10	0.06	0.46
Plataforma	**1.00**	0.03	0.06	0.07	0.07	0.13	0.33	0.30	0.11	0.06	0.12	0.10	0.06	0.18
Arquitetura	**1.00**	0.03	0.06	0.07	0.07	0.13	0.10	0.10	0.11	0.05	0.12	0.07	0.06	0.37
Domínio	**1.00**	0.03	0.06	0.07	0.07	0.13	0.10	0.10	0.11	0.05	0.12	0.07	0.06	0.43
Sprint_ID	0.00	0.45	0.49	0.37	0.45	0.48	0.49	**0.52**	0.34	**0.52**	0.47	**0.55**	0.48	0.44
US_ID	0.00	**0.54**	**0.59**	**0.69**	**0.56**	**0.60**	**0.66**	**0.62**	**0.53**	**0.69**	**0.56**	**0.60**	**0.53**	0.47
Tarefa_ID	0.00	0.47	0.48	0.40	**0.52**	**0.54**	0.46	0.49	**0.59**	0.40	**0.60**	**0.51**	**0.52**	0.35
Descritivo_da_US	**1.00**	**0.52**	**0.59**	**0.67**	**0.56**	**0.60**	**0.66**	**0.62**	**0.53**	**0.64**	**0.56**	**0.57**	**0.53**	**0.59**
Modulo	**1.00**	0.24	0.09	0.21	0.25	0.34	0.33	0.33	0.35	0.10	0.12	0.12	0.07	0.17
Operacao	**1.00**	0.45	0.37	0.41	0.45	0.48	**0.53**	**0.50**	**0.51**	0.26	0.49	0.36	0.32	0.35
Tarefa_mapeada	**1.00**	**0.53**	**0.52**	**0.64**	**0.60**	**0.71**	**0.67**	**0.66**	**0.81**	0.46	**0.67**	0.49	0.37	0.44
Tarefa_original	**1.00**	**0.91**	**1.00**	**0.89**	**0.93**	**1.00**	**1.00**	**1.00**	**0.90**	**1.00**	**1.00**	**0.95**	0.37	**0.88**
Camada	**1.00**	0.28	0.30	0.26	0.29	0.31	0.33	0.31	0.33	0.24	0.34	0.29	0.29	0.25
Linguagem	**1.00**	0.37	0.36	0.29	0.29	0.35	0.35	0.40	0.11	0.35	0.34	0.42	0.32	0.33
Framework	**1.00**	0.34	0.33	0.28	0.44	0.42	0.48	0.43	0.39	0.33	0.42	0.34	0.37	0.40
API	**1.00**	0.26	0.24	0.12	0.36	0.42	0.08	0.33	0.11	0.18	0.33	0.37	0.04	0.17
Persistencia	**1.00**	0.33	0.28	0.43	0.25	0.36	0.19	0.29	0.24	0.23	0.35	0.30	0.32	0.25
Outras_Tags	**1.00**	0.42	0.34	0.41	0.40	**0.52**	**0.55**	**0.55**	0.26	0.20	0.26	0.43	0.06	0.29
Esforco_estimado_em_horas	0.00	0.37	0.39	0.33	0.42	0.35	0.08	0.47	0.42	0.45	**0.53**	0.44	0.32	0.32
Esforco_em_horas	0.00	0.39	0.05	0.47	0.06	0.46	0.08	0.07	0.52	0.05	0.09	0.07	0.49	0.27
NFR_Tipo	0.00	0.07	0.05	0.05	0.06	0.10	0.08	0.07	0.09	0.05	0.09	0.07	0.04	0.02
NFR_Atributo	0.00	0.07	0.05	0.05	0.06	0.10	0.08	0.07	0.09	0.05	0.09	0.07	0.04	0.02
NFR_Sentença	0.00	0.03	0.05	0.05	0.06	0.10	0.08	0.07	0.09	0.05	0.09	0.07	0.04	0.01
Responsavel	0.00	0.42	0.43	0.44	0.39	0.44	0.47	**0.50**	**0.52**	0.41	0.47	0.42	0.42	0.47

Tabela 10 – Parâmetros para cálculo das métricas de avaliação do motor de inferência estatística

	VP	FN	FP	VN	Precision	Recall	F-measure	Accuracy	Informedness	Markedness
Especialista	115	19	0	929	1.00	0.86	0.92	0.98	0.86	0.98
IRI	134	0	9	920	0.94	1.00	0.97	0.99	0.99	0.94

APÊNDICE B: REVISÃO DA LITERATURA

Na Tabela 11 é apresentado um resumo da avaliação das publicações levantadas durante a revisão da literatura.

Tabela 11 – Avaliação das publicações

ID	Autores	C1	C2	C3	C4	C5	Qualidade (%)
S01	(HOGAN et al., 2012)	1	1	1	1	0.44	88.80
S02	(HU; YANG; QU, 2014)	1	1	1	1	0.29	85.74
S03	(SONG; HEFLIN, 2013)	1	1	1	1	0.00	80.00
S04	(CHEN et al., 2019)	1	1	0	1	0.71	74.10
S05	(RISTOSKI; BIZER; PAULHEIM, 2015)	1	1	0	1	0.44	68.80
S06	(HU; JIA, 2015)	1	1	0	1	0.44	68.80
S07	(PERNELLE; SAÏS; SYMEONIDOU, 2013)	1	1	0	1	0.44	68.80
S08	(NARDUCCI; PALMONARI; SEMERARO, 2017)	0	0	1	1	1.00	60.00
S09	(ALAM et al., 2017)	1	0	0	1	0.92	58.48
S10	(MEYMANDPOUR; DAVIS, 2016)	1	0	1	0	0.92	58.48
S11	(MONGIOVÌ et al., 2016)	1	0	0	1	0.92	58.48
S12	(GANGEMI et al., 2016)	1	0	0	1	0.92	58.48
S13	(ZONG et al., 2015)	1	0	0	1	0.92	58.48
S14	(LIU et al., 2018)	1	0	0	1	0.74	54.75
S15	(COLUCCI et al., 2016)	1	0	0	1	0.44	48.80
S16	(KEJRIWAL; MIRANKER, 2015)	1	0	0	1	0.44	48.80
S17	(CHENG; XU; QU, 2015)	1	0	0	1	0.44	48.80
S18	(GROZA et al., 2011)	0	1	0	1	0.43	48.68
S19	(PFAFF; KRCMAR, 2018)	0	1	0	1	0.38	47.68
S20	(GALLINUCCI et al., 2018)	1	0	0	1	0.37	47.49
S21	(DEMARTINI; DIFALLAH; CUDRÉ-MAUROUX, 2013)	1	0	0	1	0.36	47.15
S22	(SLEEMAN; FININ; JOSHI, 2015)	1	0	0	1	0.35	46.96
S23	(NGOMO, 2012)	1	0	0	1	0.31	46.27
S24	(ZHENG et al., 2015)	0	1	0	1	0.28	45.53
S25	(SHAO et al., 2016)	1	0	0	1	0.21	44.29
S26	(OLIVEIRA et al., 2019)	1	0	0	1	0.15	43.08
S27	(SALVADORI et al., 2017)	1	0	0	1	0.15	43.08
S28	(MOUNTANTONAKIS; TZITZIKAS, 2018)	1	0	0	1	0.00	40.00
S29	(SONG; ZACHAREWICZ; CHEN, 2013)	1	0	0	0	0.68	33.67
S30	(PILEGGI, 2016)	1	0	0	0	0.45	29.06