AKR Advances on Knowledge Representation Journal

O Papel da Identidade Semântica do Conceito com Vistas à Recuperação Inteligente da Informação: Do Registro do Conceito às Coordenadas Semânticas

Nilson Theobald Barbosa[1], Maria Luiza de Almeida Campos[2]

Resumo: *Discute-se o desenvolvimento de sistemas de recuperação de informação inteligentes capazes de promover a interoperabilidade semântica entre bases de dados distintas e heterogêneas. Destaca-se a importância e a viabilidade de estabelecer mecanismos que permitam determinar identidades semânticas para os conceitos presentes em sistemas de organização do conhecimento (SOC), de modo que possam ser compatibilizados e mapeados pelo seu conteúdo e significado semântico, possibilitando a criação de sistemas de recuperação da informação inteligentes entre vocabulários heterogêneos. O objetivo central consiste em permitir que agentes de software identifiquem, de maneira automatizada e precisa, relações de equivalência e proximidade entre vocabulários, superando as limitações dos localizadores tradicionais da web e transformando dados isolados em um espaço semântico navegável e integrado. A fundamentação teórica apoia-se nos métodos da Metalinguagem da Economia da Informação, de Pierre Lévy e no Registro do Conceito de Dahlberg, além de contribuições de algoritmos oriundos da Ciência da Computação e áreas correlatas, como análise de grafos e similaridade de cadeias de caracteres, a fim de mapear significados além da simples forma verbal. Adota-se uma metodologia de pesquisa bibliográfica, exploratória e qualitativa. Como resultado, apresentam-se caminhos para a utilização do registro do conceito como um método viável para nossos propósitos de interoperabilidade entre SOC, indicando perspectivas e desdobramentos para trabalhos futuros.*

Palavras-chave: *Interoperabilidade Semântica; Coordenadas Semânticas; Registro do Conceito; Recuperação Inteligente da Informação; Sistemas de Organização do Conhecimento.*

1. Introdução

O avanço das tecnologias da informação foi capaz de oferecer crescentes facilidades no armazenamento e compartilhamento de dados pela Internet e pela Web. Temos hoje recursos para criação e armazenamento de grandes bases de dados e temos redes com grande vazão de dados, disponíveis em todos os lugares, e todas estas tecnologias permitem

[1] Doutor em Ciência da Informação (UFF). Analista de Tecnologia da Informação. Universidade Federal do Rio de Janeiro (UFRJ).
E-mail: nilson@tbarbosa.org.ORCID: https://orcid.org/0000-0003-1497-313X
[2] Doutora em Ciência da Informação (UFRJ/IBICT). Docente do PPGB - UNIRIO. Universidade Federal do Estado do Rio de Janeiro (UNIRIO).
E-mail: marialuizalmeida@gmail.com. ORCID: https://orcid.org/0000-0002-9253-3706

que, além da diversidade, estes dados digitais sejam criados em volume cada vez maior e cada vez mais rápido, por diferentes grupos sociais, em diferentes contextos, usando diferentes idiomas, com diferentes compromissos ontológicos, enfim totalmente heterogêneos em todos os seus aspectos.

Estas tecnologias, que por um lado permitem e facilitam a criação e disseminação da informação, numa visão democrática e livre, têm por outro lado uma contrapartida, ou seja, a dificuldade de recuperar informação em ambientes de multivocabulários interligados, de bases de dados heterogêneos, em ambientes abertos e não controlados, tais como a Web, ou mesmo dentro de grandes empresas.

Nossa pesquisa se volta e se dedica a estudar caminhos que apontem para a compatibilização e mapeamento de conceitos que permitam a recuperação inteligente da informação em diferentes sistemas de organização do conhecimento (SOC) presentes em ambientes abertos, sem a modificação dos vocabulários utilizados e de forma automatizada por agentes de software (Barbosa & Campos, 2022). Especificamente neste artigo apresentaremos um recorte importante desta pesquisa, abordando a necessidade de estabelecimento de uma identidade dos conceitos a serem compatibilizados, para que este processo ocorra de forma semântica, ou seja, utilizando não simplesmente a forma verbal dos conceitos presentes nos SOC, mas sim seu conteúdo conceitual, sua carga semântica, em cada contexto abordado.

Para que seja possível chegar a um sistema deste tipo é preciso romper a barreira do simples apontamento fornecido pelos *Uniform Resource Locators* (URL), que são a forma de identificar recursos na Web de hoje e que servem em seu propósito simples de determinar a sua localização, mas por serem totalmente opacos, ou seja, não representarem de forma transparente o seu significado, não são capazes de, por si só, participarem de uma recuperação semântica. A utilização de recursos computacionais voltados para manipulação e avaliação de caracteres e tratamento de hierarquias e grafos é o caminho que permite a criação de medidas de apontamentos entre os conceitos de forma transparente e semântica, através do estabelecimento de suas medidas semânticas de compatibilidade. Estas medidas semânticas, segundo a definição de Harispe *et al.* (2015, p.13) são "ferramentas matemáticas utilizadas para estimar a força da relação semântica entre unidades de linguagem, conceitos ou instâncias, através de uma descrição numérica obtida de acordo com a comparação de informações que sustentam seu significado".

Com o propósito de atingir nossos objetivos, de interligar semanticamente vocabulários e bases de dados heterogêneas, as informações extraídas dos SOC para cada conceito devem ser capazes de estabelecer uma identificação semântica que pode ser objeto de processos computacionais que permitam definir relacionamentos e apontamentos entre conceitos de diferentes SOC, heterogêneos em suas estruturas e linguagens, mas possíveis de serem compatibilizados em um sistema de recuperação da informação inteligente.

Este é um ponto central de nossa pesquisa que, ao utilizar as propostas do registro do conceito de Dahlberg (1981) e visando uma identidade conceitual pressuposta por Lévy (2014), como veremos a seguir, permite que algoritmos computacionais atuem sobre os

conceitos, estabelecendo medidas numéricas de compatibilidade semântica entre eles. Estas relações e apontamentos de conceitos entre diferentes SOC são a base de um possível sistema inteligente de recuperação da informação que permita a um usuário extrair informações de qualidade entre bases de dados heterogêneas.

Na construção deste artigo utilizamos uma metodologia de pesquisa bibliográfica, exploratória e qualitativa, que nos permitisse partir de estudos já realizados, importantes para o surgimento de novos caminhos, utilizando-se de fontes bibliográficas. Além disso, seu caráter exploratório tem por finalidade gerar maiores informações sobre o tema, possibilitando a construção de definições e facilitando sua delimitação. Por fim, seguimos o caminho de uma abordagem qualitativa que supõe um desenvolvimento interpretativo de dados coletados nas fontes bibliográficas consultadas.

Assim, este artigo está organizado da seguinte forma: após esta introdução, na seção 2 discutimos a proposta de coordenadas semânticas de Pierre Lévy; na seção 3 apresentamos o registro do conceito de Ingetraut Dahlberg; na seção 4 abordamos a importância e a metodologia de elaboração de uma identidade de cada conceito, de modo a subsidiar um sistema de recuperação inteligente que possibilite a interoperabilidade entre SOCs heterogêneos e abertos; na seção 5 descrevemos técnicas computacionais capazes de viabilizar o processamento de informações que aproximem conceitos com expressões verbais distintas, mas com carga semântica equivalentes; na seção 6 discutimos a aplicação dessas técnicas associadas ao registro de conceitos de Dahlberg visando o estabelecimento de apontamentos inteligentes entre conceitos de diferentes SOCs que possam compor um sistema de recuperação inteligente da informação; por fim, na seção 7 apresentamos nossas considerações finais.

2. As Coordenadas Semânticas de Pierre Lévy

Entre as contribuições teóricas importantes que incluímos em nosso estudo para endereçar e justificar a necessidade de estabelecimento de identidades semânticas dos conceitos, temos as coordenadas semânticas propostas por Pierre Lévy, que vem se dedicando a explicitar uma construção teórica que ele denomina de Metalinguagem da Economia da Informação (IEML) (Lévy, 2014). O autor argumenta que a sua principal hipótese para propor tal metalinguagem é a de que ainda não inventamos sistemas simbólicos que se encaixam no novo meio digital que possam atender plenamente as necessidades atuais de recuperação da informação.

Ao propor essa construção, Lévy define a IEML como: 1. uma linguagem artificial que se traduz automaticamente em línguas naturais; 2. uma linguagem de metadados destinada à marcação semântica colaborativa de dados digitais; 3. uma nova camada de endereçamento do meio digital (endereçamento conceitual), que resolve o problema da interoperabilidade semântica; 4. uma linguagem de programação especializada no design de redes semânticas; 5. um sistema de coordenadas semânticas da mente (a esfera

semântica), que permite a modelagem computacional da cognição humana e a auto-observação das inteligências coletivas (Lévy, 2014).

Quanto aos seus propósitos, a IEML pode ser entendida como uma metalinguagem artificial que simultaneamente atua como uma linguagem de programação, uma linguagem de intermediação e um sistema de coordenadas semânticas. Ao categorizar dados na IEML, a metalinguagem computa suas relações e distâncias semânticas. Lévy explica que a IEML pode resolver os seguintes problemas:

1. Descompartimentalização de tags, folksonomias, taxonomias, ontologias e idiomas;
2. Pesquisa semântica, computação automática e visualização de relações semânticas e distâncias entre os dados;
3. Devolução aos usuários das informações que produzem, possibilitando a inteligência coletiva reflexiva (Lévy, 2016).

Num caminho que interessa à nossa discussão sobre a criação de identificadores únicos com significado semântico para compatibilização de conceitos, verificamos que a IEML é uma linguagem formal que calcula, gera e reconhece automaticamente uma infinidade de conceitos e seus relacionamentos semânticos. Em suma, "a IEML não é uma ontologia universal, é uma linguagem com semântica calculável que pode expressar qualquer ontologia" (Lévy, 2019, p. 25).

A proposta integral de Lévy traz ideias que são desafiadoras para implementação devido à sua complexidade e abrangência. Contudo, ela oferece abordagens metodológicas para o desenvolvimento de mecanismos que facilitam a transição entre diversas ontologias e vocabulários, que são de especial interesse para nossa investigação. Uma dessas ideias, que suporta todo o conjunto da proposta de Lévy e embasa o que perseguimos neste artigo para sua implementação no mundo real, é a criação dos identificadores singulares denominados *Uniform Semantic Locators* (USL). As operações computacionais efetuadas nas séries de sequências, que são os USL, correspondem também a operações sobre os significados que essas séries simbolizam, isto é, os conceitos. "O ponto crucial a ser compreendido é que qualquer trajetória no espaço hipertextual das ligações entre USL pode ser descrita por uma função, e que tal função pode possuir relevância semântica" (Lévy, 2014, p.478).

A proposta de Lévy nos conduz à compreensão de que os conceitos e suas representações em um sistema de comutação para recuperação de informações entre várias linguagens podem ser recuperados por meio de métodos de mapeamento. Observamos que, em sua proposta, é essencial que esses identificadores únicos apresentem descritores que expressem seu significado semântico, ou seja, os USL, ao contrário dos existentes URL, e suas interligações formariam um espaço de coordenadas semânticas voltado para a recuperação da informação.

Ao abordar e propor a criação do espaço de coordenadas semânticas, Pierre Lévy (2014) ressalta a necessidade imperativa de estabelecer um novo sistema de identificação e endereçamento dos conceitos, considerando a progressiva evolução do meio digital. Essa evolução inicia na década de 1950, quando vivemos a camada da codificação digital básica, da manipulação simples de símbolos, linguagens de programação e endereçamento de bits, de forma interna aos computadores. Nos anos 1970-1980, passamos às conexões entre os autômatos (computadores), usando neste processo protocolos de internet baseados em TCP/IP para identificar e ligar os atores participantes, e surgem então diversos tipos de agentes, tais como os roteadores. Por volta de 1995, com a Web, temos as conexões entre os dados, e a sua identificação realizada através dos *Uniform Resource Locators* (URL). Para a concretização da sua proposta de esfera semântica, Lévy, propõe a criação dos USL, cuja função é a identificação e endereçamento dos conceitos. Esta proposta claramente não se coloca como uma substituição das camadas anteriores, e não prescinde delas, da mesma forma que a camada da web não substituiu as suas camadas prévias, pois, considerando as tecnologias atuais, será necessário endereçar dados no meio digital, em seus diversos níveis, usando protocolos de internet e apontamentos por URL. Nesse caso, se acrescenta uma nova camada de codificação, que permitirá interpretar e utilizar conceitos melhor do que se faz com dados da web e seus URL (Barbosa, 2021).

A IEML constitui uma linguagem formal estruturada em camadas e gerada a partir de seis primitivas semânticas: S (sinal), B (ser), T (coisa), U (virtual), A (atual) e E (vazio) (Lévy, 2023). Suas expressões sintaticamente válidas podem ser geradas e reconhecidas por um autômato finito. No contexto do IEML, o número de expressões válidas de significação distinta, ou seja, conceitos únicos, é finito, embora extremamente vasto, onde a quantidade de identificadores distintos excede o número de partículas elementares do universo físico conhecido.

Cada texto na linguagem, ou, conforme denominado por Lévy, cada unidade semântica (o USL), pode ser considerado um nó na esfera semântica. Os autômatos que compõem o sistema completo do IEML são capazes de traçar circuitos semânticos entre esses USL. Este autômato semântico conecta todos os nós da esfera IEML por meio de dois tipos de ligações: as ligações paradigmáticas, que conectam expressões considerando seus aspectos conceituais, e as ligações sintagmáticas, que conectam as expressões considerando seus aspectos de enunciados (Lévy, 2014, 2019).

Ao considerarmos estas proposições e métodos de Lévy, fomos buscar nos estudos seminais da Ciência da Informação propostas e métodos capazes de participarem de um processo de implementação de um sistema de recuperação da informação que pudesse implementar este ambiente de coordenadas semânticas. Num ambiente deste tipo cada conceito tem sua própria identificação e esta identificação representa sua carga semântica e, com isso, pode ser o objeto deste processo de recuperação inteligente e semântica da informação. Isso se alinha ao trabalho de Dahlberg (1981), que buscou descrever os

conceitos de um sistema de indexação por meio de uma entidade denominada pela autora como o Registro do Conceito, visando o tratamento e a recuperação da informação, que será apresentado na próxima seção.

3. O Registro do Conceito de Dahlberg

O registro do conceito é um dispositivo que se coloca no âmbito do método apresentado por Dahlberg (1981) denominado Matriz de Compatibilização Conceitual. Tal método propõe um mapeamento da potencialidade semântica dos vocabulários, fornecendo esta análise de compatibilidade entre as linguagens em questão sob os pontos de vista semântico e estrutural.

A primeira etapa para a elaboração de uma matriz de compatibilidade é o que a autora denomina como um procedimento de casamento verbal ou linguístico dos termos, que pode ser automatizado e registrado em uma matriz preliminar.

A segunda etapa é avaliar o percentual do possível casamento verbal e semântico entre os termos. Essa avaliação permite identificar o grau de correspondência conceitual encontrado a partir da compatibilidade no plano linguístico. Entretanto, verifica-se também que o casamento linguístico não é suficiente para detectar ocorrências entre nomenclaturas diferentes para termos com o mesmo significado, homônimos ou ainda termos linguisticamente semelhantes, mas que estão especificados em diferentes cadeias de conceitos. Dessa forma, a matriz preliminar obtida deve ser complementada, a partir dessa segunda etapa, por meio da análise dos conceitos, para que uma correspondência semântica possa ser estabelecida. É nesse momento que se estabelece uma terceira etapa.

A terceira etapa é a elaboração de um dispositivo que ela denominou de registro do conceito para que seja possível inferir o nível de compatibilidade conceitual, ou seja: a coincidência conceitual, a correspondência conceitual e a correlação conceitual.

Os tipos de nível de compatibilidade entre linguagens, segundo Dahlberg, compreendem três fases: (i) a coincidência conceitual – dois conceitos combinam suas características – grau de equivalência; (ii) correspondência conceitual – dois conceitos combinam a maior parte de suas características – similaridade; (iii) correlação conceitual – dois conceitos são correlacionados através de símbolos matemáticos, estabelecendo uma medida de correlação, quando possuem diferentes níveis de detalhe, ou quando a relação entre eles não é de semelhança, sendo similares; e correlação conceitual, dois conceitos são correlacionados por meio de símbolos matemáticos. O tipo de correlação pode ser indicado por símbolos, por exemplo: “menor” e “maior” para indicar diferentes níveis de detalhamento, “C” para indicar que um conceito em uma linguagem equivale a uma combinação de conceitos na outra (Dahlberg, 1983).

Como resultado dessa análise conceitual complementar, obtém-se a matriz de compatibilidade final, denominada por Dahlberg de matriz de compatibilidade conceitual, que estabelece, além da correspondência dos conceitos, uma medida de compatibilidade,

e o tipo de correspondência (“maior”, ‘menor” etc.), conforme citado anteriormente. A matriz de compatibilização conceitual funciona como um dispositivo semântico no qual se entra com um descritor de um dado vocabulário e obtém-se o descritor correspondente nos outros vocabulários.

Ao entendermos a necessidade de buscar a criação de uma identificação única para os conceitos de determinado ambiente e que esta identificação deva ser semanticamente representativa da significação (ou da semântica) de cada conceito, encontramos no trabalho de Dahlberg (1981) o caminho que endereça este problema. Dahlberg, ao propor o seu registro do conceito, já apresentava preocupação nesse sentido, apesar de sua implementação se dar em matrizes de correlação manuais, em especial devido ainda aos parcos recursos computacionais existentes em sua época de publicação.

Dahlberg propõe métodos para estabelecer comparações verbais e conceituais entre linguagens de indexação através de seus descritores, bem como para criar matrizes de compatibilização. No primeiro caso, a autora define o estabelecimento de uma matriz de comparação voltada para as formas verbais dos termos. Quando duas ou mais linguagens de indexação são comparadas em relação à sua compatibilidade, é necessário obter uma visão inicial da sobreposição real de classes e descritores que correspondem aos termos verbalmente. Nesse caso, o conteúdo de cada descritor deve ser registrado e seus conceitos analisados num registro conceitual (quadro 1) com as seguintes entradas: Código atribuído e Nome do conceito (1); Fonte ou Notação (2); Próximo conceito abrangente (3); Conceito mais alto na hierarquia (4); Indicação do nível hierárquico do conceito (5); Número de subconceitos (6); Categoria do conceito (7); Definição do conceito (8); Outros nomes do conceito (9); Abreviações (10); e Comentários (11) (Dahlberg, 1981).

Quadro 1. Registro do conceito

00	Código do Conceito
01	Nome do conceito
02	Notação do conceito (quando for o caso de se compatibilizar Sistemas de Classificação) ou Fonte do conceito (nome do Sistema Ordenado ou um código definido para indicá-lo)
03	Denominação do próximo conceito mais amplo
04	Denominação do conceito mais alto na hierarquia
05	Indicação do nível hierárquico do conceito (formando uma cadeia de conceitos): Nível mais alto, Próximo nível mais baixo, Terceiro nível etc.
06	Número de subconceitos
07	Categoria da forma do conceito (O) Objeto, entidade (P) Processo, atividade (Q) Quantidade, qualidade (R) Relação (S) Conceito relacionado a espaço (T) Conceito relacionado a tempo (W) Campo de assunto ou disciplina
08	Definição do conceito e com indicação das fontes

09	Outros nomes do conceito (sinônimos)
10	Abreviações do Termo
11	Comentários relativos aos conceitos correspondentes em outros Sistemas Ordenados

Fonte: Elaborado pelos autores, a partir de Dahlberg (1981)

O estudo e compreensão do registro do conceito de Dahlberg, com a definição de suas aplicações nos processos de recuperação da informação, informatizados e automáticos, através do estabelecimento de medidas semânticas de compatibilidade entre conceitos de diferentes sistemas de organização do conhecimento, é o resultado esperado desta faceta de nossa pesquisa, conforme explicitamos na seção 4 a seguir.

4. Elaboração de uma Identidade Semântica para Conceitos em SOC Visando a Interoperabilidade

Apesar da necessidade humana contemporânea de construir uma inteligência coletiva e distribuída, acessível e aberta, o que observamos no cenário real da Web é uma rede fragmentada por diversos produtos e serviços concorrentes, múltiplas línguas e uma grande diversidade sistemas de organização do conhecimento, elaborados e desenvolvidos com diferentes tecnologias. Esta situação se manifesta em escala global quando discutimos a recuperação da informação na Web, bem como internamente dentro das corporações, onde, apesar dos esforços de compatibilização, as bases de dados e os sistemas de organização do conhecimento de modo geral ainda não conseguem se comunicar de maneira semântica, compreensível e útil, mesmo em bases de dados locais e de uso restrito.

É neste contexto que emerge o desafio contemporâneo da busca da interoperabilidade com vistas à recuperação da informação. A implementação de processos de recuperação da informação que possam relacionar dados semanticamente, promovendo a interoperabilidade entre bases heterogêneas, preferencialmente de maneira automática e com mínima intervenção manual, sem a necessidade de convergir para um único formato, se coloca como um desafio significativo para os dias de hoje.

Esses processos, definidos aqui como recuperação inteligente da informação, são essenciais para a construção dos alicerces de uma nova web. Suas premissas fundamentais incluem: permitir que a informação seja recuperada a partir de bases de dados indexadas por vocabulários heterogêneos e possibilitar que agentes de software participem do processo de recuperação da informação de forma automática, em conjunto com a ação interativa de seres humanos.

Ao tentar identificar a carga semântica dos conceitos em SOC tais como taxonomias e tesauros, com o propósito de compatibilizá-los, precisamos conseguir extrair estas informações da estrutura presente nestes vocabulários. Uma vez que as únicas informações capazes identificar semanticamente os conceitos, seja em taxonomias com suas relações de superordenação ou de subordinação, ou as relações internas

adicionais existentes nos tesauros, são estas formas verbais, é com elas que precisamos trabalhar.

Nosso caminho, no escopo deste artigo, tomando como base a proposta de Lévy de criar uma linguagem de universal de recuperação da informação – na qual cada conceito é representado por um identificador único - é estabelecermos que a elaboração de uma identidade para cada conceito é essencial para definir sua individualidade e, posteriormente, suas relações semânticas com outros conceitos nos diferentes vocabulários envolvidos no processo de compatibilização. Para a constituição desta identidade vamos lançar mão da proposta de registro de conceito de Dahlberg, conforme definido no quadro 1 acima. As informações coletadas e armazenadas neste registro são capazes de proporcionar recursos que permitiriam agentes de software realizarem comparações textuais entre os vocabulários, ou seja, ao considerar estas informações registradas, o conceito não é mais identificado somente pelo conjunto de caracteres e palavras que o representa inicialmente, que compõe sua expressão verbal, mas sim pelo conjunto de informações que pode ser extraído dos SOC, de forma automática, ultrapassando a fronteira da interoperabilidade sintática entre o nome do conceito e partindo para uma tentativa de interoperabilidade semântica.

Antes de prosseguirmos, é importante, neste momento, definir o propósito de estudar métodos para definir identidades semânticas para os conceitos, de forma a identificá-los de forma única. Ao estabelecer estas identidades podemos caminhar para o estabelecimento de processos automáticos e agentes de software capazes de comparar conceitos de fontes diferentes e conseguir aplicar medidas de similaridade semântica que definam quanto numericamente um conceito é similar a outro conceito, entre sistemas de organização do conhecimento heterogêneos.

A similaridade semântica é uma métrica utilizada para medir a proximidade entre documentos ou termos, levando em consideração a semelhança de seus significados ou conteúdos semânticos, em vez de apenas suas características lexicográficas. Trata-se de ferramentas matemáticas empregadas para avaliar a intensidade da relação semântica entre unidades de linguagem, conceitos ou instâncias, por meio de uma descrição numérica obtida ao comparar as informações que sustentam seus significados ou descrevem suas naturezas (Feng *et al.*, 2017; Harispe *et al.*, 2015).

As medidas de similaridade semânticas são amplamente empregadas na atualidade para a comparação de entidades semânticas, tais como unidades de linguagem, conceitos ou mesmo instâncias semanticamente caracterizadas, com base nas informações que sustentam seu significado. Fundamentam-se na análise de corpus de textos estruturados ou não-estruturados, bem como em sistemas de organização do conhecimento, dos quais se podem extrair evidências semânticas. A literatura da área aponta que a noção de medida semântica não se enquadra estritamente na definição matemática rigorosa de medida. Em vez disso, deve ser compreendida como qualquer

ferramenta teórica, função matemática, algoritmo ou abordagem que permita a comparação de entidades semânticas com base em evidências semânticas.

De modo geral, essas medidas são utilizadas para estimar o grau de similaridade semântica entre entidades semânticas por meio de um valor numérico. Nesse contexto, embora exista uma vasta diversidade de medidas para estimar a similaridade ou a distância entre objetos matemáticos específicos (por exemplo, vetores, matrizes, grafos, conjuntos, conjuntos nebulosos), estruturas de dados (por exemplo, listas, árvores) e tipos de dados (por exemplo, números, cadeias de caracteres, datas), a principal particularidade das medidas semânticas em comparação com as funções tradicionais de similaridade ou distância reside em dois aspectos: (i) são dedicadas à comparação de entidades semânticas e (ii) baseiam-se na análise de onde as evidências semânticas podem ser extraídas.

Essas evidências semânticas têm a finalidade de caracterizar, direta ou indiretamente, o significado dos elementos comparados. Por exemplo, as medidas utilizadas para comparar duas palavras de acordo com suas sequências de caracteres não podem ser consideradas semânticas, uma vez que levam em conta apenas os caracteres das palavras e sua ordem, sem considerar seus significados. De fato, de acordo com tais medidas sintáticas, as palavras "carro" e "veículo" seriam consideradas distantes, apesar de sua semântica intimamente relacionada. As medidas semânticas permitem superar a limitação dessas medidas sintáticas, comparando entidades semânticas com base em sua semântica. Para isso, as medidas semânticas utilizam a análise de dois tipos amplos de proxies semânticos[3]: corpus de textos e ontologias. Os proxies semânticos são empregados para extrair evidências semânticas que serão posteriormente utilizadas pelas medidas semânticas para apoiar a comparação de unidades de linguagem, conceitos ou instâncias (Harispe *et al.,* 2015).

Podemos afirmar que o objetivo a ser alcançado com as medidas semânticas reside na sua capacidade de evidenciar o nível de interação semântica entre elementos com diferentes níveis de semântica, tais como palavras e conceitos, com base em seus significados. Por exemplo, é plausível que as palavras "escola" e "colégio" estejam semanticamente mais relacionadas do que "escola" e "árvore". Contudo, podemos medir essa relação? Alguns métodos e algoritmos, discutidos a seguir neste artigo, propõem etapas para atingir tal objetivo, buscando extrair esse conhecimento das relações explicitadas em sistemas de organização do conhecimento e emular a capacidade humana de estabelecer o grau de relação entre diferentes 'entidades' de acordo com as evidências semânticas disponíveis. A seguir apresentamos um resumo das técnicas computacionais que podem ser aplicadas sobre os registros de conceitos extraídos dos SOC, de forma a estabelecer as medidas semânticas de similaridade.

[3] Um proxy semântico é qualquer fonte de informação da qual pode ser extraída indicações da semântica de seus elementos, que por sua vez serão usados em uma medida semântica.

5. Técnicas Computacionais

As técnicas computacionais apresentadas e discutidas aqui referem-se a algoritmos e agentes de software existentes ou concebidos especificamente para estabelecer medidas de similaridade semântica entre conceitos presentes em diferentes repositórios. Tais recursos funcionam como ferramentas matemáticas que estimam a força da relação entre unidades de linguagem, conceitos ou instâncias, traduzindo o significado em descrições numéricas baseadas na comparação de informações que sustentam sua natureza. O objetivo é permitir que computadores processem informações de forma transparente, rompendo a barreira da opacidade dos identificadores tradicionais da Web, que apenas localizam recursos sem representar seu conteúdo.

Esses processos computacionais buscam emular a capacidade humana de identificar o grau de relação entre diferentes entidades de acordo com evidências extraídas de sistemas de organização do conhecimento. Ao contrário de buscas sintáticas tradicionais, que se limitam a comparar sequências de caracteres, estas técnicas utilizam proxies semânticos para extrair evidências que caracterizam, direta ou indiretamente, o real significado dos elementos comparados. Dessa forma, torna-se possível aproximar conceitos que podem possuir expressões verbais distintas, mas que compartilham uma carga semântica similar.

Nesse sentido, os processos computacionais a serem citados abaixo, agentes de software, ou simplesmente algoritmos, que podem fornecer estas medidas, podem ser divididos em (a) técnicas de elementos, baseadas em cadeias de caracteres, em linguagens ou em restrições, que denominamos de Técnicas de Nível de Elemento na Medição da Similaridade Semântica e (b) técnicas de estrutura, baseadas em taxonomias, em grafos, em instâncias ou em modelos que denominamos de Técnicas de Nível de Estrutura na Medição de Similaridade Semântica. Todas estas, e os algoritmos de semântica distribucional e de Word Embedding, para manipulação de textos, podem ser aplicados para análise dos SOCs e dos registros de conceito apurados, onde não atuarão simplesmente sobre a forma verbal do conceito e sim sobre uma identidade extraída do sistema de organização do conhecimento onde reside o conceito, capaz de transmitir o significado daquele conceito, e estabelecimento de suas medidas de similaridade entre todos os outros conceitos de nosso proposto espaço semântico de recuperação da informação. Apresentamos aqui apenas uma amostra. A relação detalhada de todas estas técnicas que tornam palpável e viabilizam nossa pesquisa podem ser obtidas em Barbosa (2021) e Angermann e Ramzan (2017).

5.1 Técnicas de Nível de Elemento na Medição de Similaridade Semântica

As técnicas de nível de elemento utilizam os valores literais dos conceitos e/ou suas propriedades para medir a similaridade semântica. Existem cinco categorias principais dessas técnicas: baseadas em recursos formais e informais, em cadeias de caracteres (strings), em linguagem e em restrições.

5.1.1 Técnicas Baseadas em Recursos Formais e Informais

As técnicas baseadas em recursos formais recorrem a conhecimentos fortemente estruturados, como taxonomias de alto nível específicas para um domínio. Exemplos incluem taxonomias padronizadas que representam grupos de diferentes domínios ou taxonomias mais gerais e abrangentes dentro de um mesmo domínio. Em contraste, as técnicas baseadas em recursos informais utilizam recursos não padronizados, como diretórios de índices estruturados em um nível superior às taxonomias a serem harmonizadas. Em ambos os casos, os elementos das taxonomias são comparados e mapeados aos elementos das taxonomias globais.

5.1.2 Técnicas Baseadas em Cadeias de Caracteres

Estas técnicas identificam correspondências comparando e igualando cadeias de caracteres. Conforme Cheatham e Hitzler (2013), a similaridade pode ser calculada de duas maneiras principais: similaridade de nome e similaridade de descrição.

A similaridade de nome avalia quão semelhante uma palavra ou grupo de palavras é a outra(s). Algumas medidas dessa similaridade incluem: **Distância de Levenshtein**: número mínimo de transformações (remoção, inserção ou substituição de caracteres) para converter uma cadeia de caracteres em outra (Levenshtein, 1966); **Distância de Bailey (Euclidiana)**: comprimento da conexão necessária para combinar pontos no espaço euclidiano (Bailey, 2004); **Processo de Hellinger**: utiliza cálculos probabilísticos (Hellinger, 2009); **Distância de Hamming**: mede diferenças entre cadeias de mesmo comprimento (Hamming, 1950; Tanenbaum, 2003); **Distância de Lin**: calcula a probabilidade de uma string ocorrer dentro de um termo (Kernighan & Lin, 1970); **Método de Wu e Palmer**: classifica termos conforme sua profundidade em um corpo de texto (Wu & Palmer, 1994).A similaridade de descrição compara termos compostos com outras sequências, utilizando medidas como: **Distância Jaccard**: mede a semelhança entre dois conjuntos de strings; **Similaridade Cosine**: trata sequências como vetores para compará-las; **TF-IDF**: considera a importância de um termo com base em sua ocorrência em um documento (Jones, 1972; Tan *et al.*, 2005).

5.1.3 Técnicas Baseadas em Linguagem

Estas técnicas são frequentemente usadas com as baseadas em cadeias de caracteres e são apoiadas pelo conhecimento do domínio para analisar o contexto dos conceitos comparados. As principais categorias incluem: **Lematização e Morfologia**: agrupam diferentes formas de inflexão de uma palavra; **Tokenização**: segmenta texto em palavras, frases ou símbolos significativos; **Eliminação de Tokens Supérfluos**: remove elementos considerados desnecessários, como stop-words; **Léxicos ou Tradutores**: traduzem entre idiomas usando tradutores automáticos, como Microsoft Bing e Google Tradutor; **Análise de Similaridade**: usa bases de dados como a WordNet para avaliar a similaridade semântica entre conceitos; **Desambiguação de Sentidos**: analisa o sentido da sentença no contexto, identificando o token mais relevante para comparação.

5.1.4 Técnicas Baseadas em Restrições

Essas técnicas analisam a estrutura interna do sistema de organização do conhecimento utilizado, frequentemente em conjunto com outros métodos, para superar a heterogeneidade conceitual. Dividem-se em: **Similaridade por Tipo de Atributos**: avalia a similaridade semântica entre conceitos com nomes diferentes, mas descrições similares em diferentes atributos. Por exemplo, "carro de passeio" e "automóvel" podem ser considerados semanticamente similares se compartilharem atributos como número de portas, espaço na mala e número de bancos; **Propriedades-Chave**: utilizadas para descrever conceitos dentro de uma taxonomia. Quando estruturadas conforme um ponto de vista correspondente, as taxonomias podem ser consideradas similares como um todo (Angermann & Ramzan, 2017).

5.2 Técnicas de Nível de Estrutura na Medição de Similaridade Semântica

Quatro técnicas principais foram destacadas no nível de estrutura: baseadas em taxonomia, em grafos, em instância e em modelo (Angermann & Ramzan, 2017).

5.2.1 Técnicas Baseadas em Taxonomia

Essas técnicas exploram os subconceitos (especialização) e superconceitos (generalização) em uma taxonomia, também conhecidos como relações do tipo é-um. As taxonomias podem diferir no número total de conceitos e nas quantidades de relações utilizadas. A análise de similaridade entre dois conceitos em diferentes estruturas avalia seus subconceitos e superconceitos. Quanto menos diferentes essas estruturas forem, maior será a similaridade semântica entre eles.

5.2.2 Técnicas Baseadas em Grafos

Nas técnicas baseadas em grafos, uma taxonomia é considerada um grafo identificado. Ao contrário das técnicas baseadas em taxonomia, as relações de paridade, ou irmandade, também são consideradas ao comparar conjuntos e subconjuntos e a distância entre cada um. São usadas técnicas matemáticas de análise de grafos, tais como homomorfismo, similaridade de caminhos, similaridade de filhos e similaridade de folhas.

5.2.3 Técnicas Baseadas em Instância

As técnicas baseadas em instância focam na similaridade entre conceitos com base em suas instâncias. Essa similaridade depende de dois conjuntos a serem comparados, definindo que conceitos similares devem ter instâncias similares. Embora nem todos os sistemas de organização do conhecimento apresentem instâncias em suas representações, esses objetos, quando presentes, são úteis para comparação e estabelecimento de mapeamentos semânticos entre conceitos.

5.2.4 Técnicas Baseadas em Modelos

As técnicas baseadas em modelos são de uso limitado na literatura e apresentam descrição dispersa e pouco densa. Essas técnicas utilizam lógicas de descrição para superar a heterogeneidade da taxonomia. Solucionadores de satisfação determinam se existe uma interpretação que satisfaça um dado operador booleano, que pode ser verdadeiro ou falso. O Raciocínio de Lógicas de Descrição, por sua vez, é uma família de linguagens formais de representação do conhecimento, capaz de inferir consequências lógicas de um conjunto de entidades (Angermann & Ramzan, 2017).

6. A aplicação das técnicas e os registros dos conceitos

Com a aplicação das técnicas computacionais de manipulação de caracteres sobre os registros de conceito extraídos e armazenados em repositórios de dados adequados e realizando os processos de alinhamento e mapeamento sobre estes registros de conceito e não somente sobre os nomes ou tags dos conceitos, caminhamos para o estabelecimento de apontamentos inteligentes entre conceitos de sistemas de organização do conhecimento diferentes e heterogêneos que possam compor um sistema de recuperação inteligente da informação (Barbosa, 2021).

A elaboração deste sistema é fundamentada na extração de registros de conceito e na execução de mapeamentos transversais, visando a consolidação de uma arquitetura de recuperação inteligente da informação. Este processo não se limita à simples identificação de equivalências, mas propõe a criação de uma identidade digital para cada conceito - o registro de conceito - que armazena informações semânticas extraídas diretamente dos Sistemas de Organização do Conhecimento (SOC) participantes. Através dessa abordagem, o conceito deixa de ser um rótulo puramente sintático e passa a ser representado por um objeto complexo que encapsula sua expressão verbal, notação, hierarquia (termos genéricos e específicos), relações associativas e definições extraídas de notas de escopo.

A operacionalização deste sistema exige que agentes de software realizem o processamento automatizado das potencialidades semânticas de cada termo, mantendo os vocabulários originais íntegros e inalterados. O fluxo de trabalho inicia-se pela identificação de identidades sintáticas exatas ou parciais (lematização, morfologia e tokenização), servindo como um filtro preliminar de candidatos ao mapeamento. Contudo, para suplantar as barreiras da ambiguidade e da polissemia, o sistema deve validar essas correspondências por meio de análises estruturais de taxonomia e grafos (vide seção acima), verificando se termos de grafia igual compartilham, de fato, o mesmo contexto semântico, ou se termos mesmo de grafia diferente tem significado igual ou semelhante.

A seguir, apresentamos uma brevíssima situação hipotética, mostrando como essas técnicas podem ser aplicadas em um cenário real de compatibilização entre dois SOC.

Como dito acima, a compatibilização de SOC — como taxonomias, ontologias ou vocabulários controlados — depende justamente de identificar quando dois termos de diferentes sistemas representam o mesmo conceito, ou seja, compartilham o mesmo significado. Desta forma, considere que duas instituições diferentes mantêm seus próprios SOC para classificar documentos acadêmicos e apresentamos como exemplo (quadro 2) três termos constantes de cada SOC.

Quadro 2. Termos a serem compatibilizados de cada SOC

Termos do SOC A	Termos do SOC B
Inteligência Artificial	IA
Aprendizado de Máquina	Machine Learning
Mineração de Dados	Descoberta de Conhecimento em Bases de Dados

Fonte: elaborado pelos autores

O objetivo é descobrir quais termos de um SOC correspondem aos do outro, mesmo quando escritos de forma diferente, utilizando meios automáticos sem ação humana. O quadro 3 abaixo mostra um resumo do resultado dos procedimentos.

Inicialmente aplicando algumas técnicas de similaridade de nome:

a) Levenshtein

Comparando *"Inteligência Artificial"* e *"IA"*:

- A distância é grande, pois as cadeias são muito diferentes.
- Resultado: baixa similaridade.

Comparando *"Aprendizado de Máquina"* e *"Machine Learning"*:

- Apesar de idiomas diferentes, algumas palavras têm padrões semelhantes.
- Similaridade moderada, mas insuficiente para afirmar equivalência.

b) Hamming

Só funciona para cadeias do mesmo tamanho. Pode ser útil para comparar abreviações, por exemplo:

- *"IA"* x *"AI"* → distância 2
- Indica que são diferentes, mas próximas.

c) Wu & Palmer

Usado quando há hierarquias semânticas. Se ambos os SOC tiverem estrutura hierárquica, por exemplo:

- SOC A: "Inteligência Artificial" → neste SOC é subclasse de "Computação"

- SOC B: “IA” → neste SOC é subclasse de “Tecnologias Digitais”

A técnica calcula a profundidade e o ancestral comum mais próximo. Resultado: alta similaridade, sugerindo equivalência.

A seguir aplicando algumas técnicas de similaridade de descrição, considerando que cada termo tem uma descrição textual associada (uma definição ou uma nota de escopo, por exemplo):

SOC A – “Mineração de Dados”

Processo de extração de padrões úteis a partir de grandes conjuntos de dados.

SOC B – “Descoberta de Conhecimento em Bases de Dados”

Conjunto de técnicas para identificar padrões relevantes em grandes volumes de dados.

a) Jaccard

Compara conjuntos de palavras:

- Palavras em comum: *padrões*, *dados*, *grandes*, *conjuntos/volumes*
- Similaridade alta → provável equivalência.

b) Cosine Similarity

Transforma as descrições em vetores:

- Termos como *padrões*, *dados*, *extração*, *identificar* aparecem com pesos semelhantes.
- Similaridade muito alta.

c) TF-IDF

Dá mais peso a termos distintivos:

- “padrões”, “dados”, “extração”, “descoberta” aparecem com alta relevância.
- As descrições convergem semanticamente.

Conclusão: “Mineração de Dados” (SOC A) pode ser compatibilizado com “Descoberta de Conhecimento em Bases de Dados” (SOC B).

Quadro 3. Resultado da compatibilização automatizada

Termo SOC A	Termo SOC B	Técnica decisiva	Resultado
Inteligência Artificial	IA	Wu & Palmer + TF-IDF	Equivalentes
Aprendizado de Máquina	Machine Learning	Cosine + TF-IDF	Equivalentes

Mineração de Dados	Descoberta de Conhecimento em Bases de Dados	Jaccard + Cosine	Equivalentes

Fonte: elaborado pelos autores

Mas além destas operações de compatibilização úteis e necessárias, apresentamos um diferencial metodológico central nesta proposta que é a transição de um mapeamento booleano (binário) entre equivalentes para o cálculo de distâncias semânticas.

Como vimos na discussão acima, termos e formas verbais textuais podem ser sintaticamente similares mas semanticamente dissimilares e, da mesma forma, podem ser sintaticamente dissimilares mas semanticamente similares (Wei *et al.*, 2025), sendo necessário avançar no estabelecimento nas medidas de similaridade semântica entre termos.

Nesse sentido, utilizando as classes de algoritmos descritos acima, desde semântica distribucional até word embedding e similares, aplicados aos termos, definições e hierarquia dos conceitos. O sistema proposto deve ser capaz de determinar não só a equivalência, mas um grau de similaridade quantificável em um intervalo entre 0 e 1 entre cada conceito e todos os outros. O resultado é um espaço multidimensional e navegacional onde os conceitos estão interconectados e interapontados por pesos de similaridade, permitindo que conceitos com expressões verbais totalmente distintas sejam reconhecidos como semanticamente próximos caso possuam estruturas relacionais análogas.

A integração dessas identidades semânticas em um repositório centralizado de registros, identificados univocamente por URIs (*Uniform Resource Identifiers*), permite a manutenção contínua e dinâmica do espaço semântico. À medida que novos SOC são integrados ou os existentes sofrem atualizações, o agente de software reprocessa as conexões, garantindo que o sistema de recuperação de informações (SRI) opere sobre uma base de conhecimento sempre atualizada. Este ambiente flexível e aberto suporta a inclusão de novas fontes e o enriquecimento progressivo da malha de conceitos interligados.

Portanto, para chegar a estes propósitos, elencamos as ações práticas a serem implementadas:

i) analisar automaticamente os SOC e extrair a inteligência presente em cada um, seja um tesauro ou mesmo uma simples taxonomia, gerando um repositório semântico das identidades de cada conceito, como apresentado no Quadro1 deste artigo;

ii) para cada um dos registros gerados, identificar registros de conceitos nos outros vocabulários participantes que sejam representados por expressões verbais consideradas semelhantes pela aplicação exaustiva das técnicas de nível de elemento, tais como tokenização, lematização, identificação de plurais, ordem dos termos invertida, extração de hifens e outras similares;

iii) para cada um dos registros gerados, identificar registros de conceitos mesmo com expressão verbal diferente nos outros vocabulários, mas que tenham semelhança em sua estrutura, usando as mesmas técnicas do item ii, em seus termos genéricos, termos específicos e termos associados;

iv) a partir daí aplicar as técnicas de análise de taxonomia e grafos, que permitem identificar similaridades pela utilização da estrutura, validando ou não os conceitos de expressão verbal igual, semelhante, ou dessemelhantes descobertos;

v) extrair o significado semântico principal das notas de escopo, através das técnicas de distribuição semântica e word embedding, de forma que permita incluir este resultado nos procedimentos de correspondência e no cálculo da distância semântica entre os conceitos;

vi) para cada situação ocorrida anteriormente, estabelecer uma medida de similaridade semântica calculada que vetorize o grau de compatibilidade de cada termo com os termos descobertos que sejam possíveis de serem compatíveis;

vii) armazenar as distâncias semânticas calculadas para cada conceito apontado, nos próprios registros de conceito, utilizando os identificadores únicos (URI) para representar os conceitos mapeados.

Como produto, o que denominamos de espaço semântico (figura 1) assim constituído oferece ao usuário a possibilidade de realizar buscas complexas em bases de dados heterogêneas de forma interativa. O sistema pode apresentar os conceitos descobertos e suas respectivas distâncias semânticas, permitindo que o pesquisador defina limiares de precisão e valide as similaridades sugeridas pelo algoritmo. Dessa forma, a construção desse espaço conceitual, inspirado na proposta de esferas semânticas de Pierre Lévy, provê os alicerces tecnológicos para transformar dados dispersos em conhecimento recuperável com altos índices de precisão e revocação.

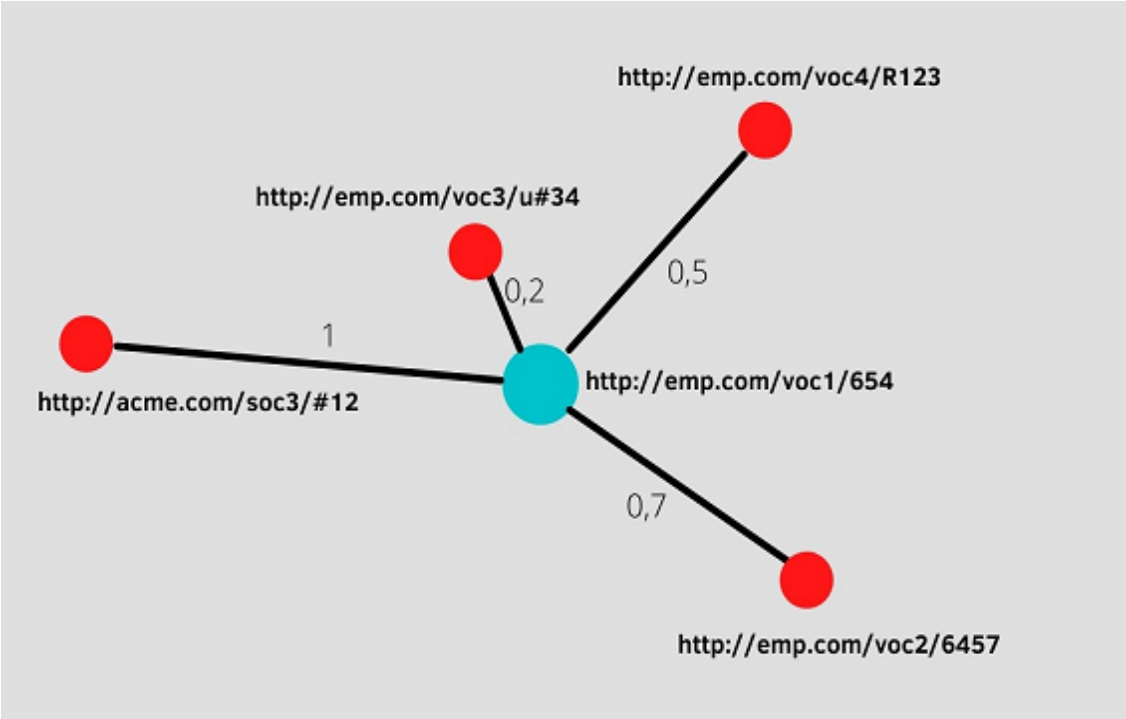

Figura 1 - Correspondências semânticas a partir de um conceito

Fonte: elaborado pelos autores

7. Considerações Finais

Neste artigo procuramos abordar um aspecto importante de nossa pesquisa, visando demonstrar um caminho possível que leve à construção de sistemas que promovam a recuperação da informação entre sistemas de organização do conhecimento heterogêneos, em ambientes tal como encontramos na Internet e na Web, ou seja em ambientes abertos e não controlados, onde não é possível intervenção humana para modificação de vocabulários e bases de dados e não há como fazer compatibilização manual entre conceitos. Temos como princípios basilares de nossa pesquisa a utilização de propostas, técnicas e métodos da Ciência da Informação que, apoiados em técnicas computacionais, possam levar à solução de problemas de interoperabilidade entre vocabulários.

Esta é uma pesquisa em andamento e como sua continuidade, o objetivo é avançar no estudo e determinação de medidas semânticas que têm como objetivo principal explicitar o nível de interação semântica entre elementos com distintos graus de significado, como palavras e conceitos, fundamentando-se em seu caráter semântico. Por exemplo, termos como "escola" e "colégio" apresentam maior proximidade semântica em comparação com "escola" e "árvore". Determinar e principalmente quantificar essa relação, no entanto, requer métodos e algoritmos capazes de medir tal proximidade com base nas evidências semânticas disponíveis. Assim, utilizando a definição de Harispe *et al.* (2015, p.13), as medidas semânticas podem ser definidas como "ferramentas

matemáticas utilizadas para estimar a força da relação semântica entre unidades de linguagem, conceitos ou instâncias, por meio de valores numéricos obtidos da comparação de informações que sustentam seu significado".

Assim, no caminho de contínuos esforços apresentados em prol da interoperabilidade de sistemas de organização do conhecimento, este estudo se propõe a avançar na quantificação das identidades semânticas entre conceitos de SOC heterogêneos, com o propósito de criar sistemas de recuperação da informação inteligentes, entre SOC em sistemas abertos e não controlados, tal como a Web, utilizando abordagens interdisciplinares entre a Ciência da Informação e seus pressupostos teóricos, e a Ciência da Computação, com algoritmos computacionais aplicados a caracteres, textos, hierarquias e grafos. Consideramos que a contribuição para o conhecimento em nossa área de estudo e pesquisa se dá com o avanço na capacidade de recuperar informação em sistemas abertos, sem intervenção manual, entre dados e sistemas heterogêneos, trazendo a possibilidade de significativos ganhos na compatibilização de entre SOC, com impacto direto na capacidade de extrair conhecimento entre sistemas de diferentes bases.

Uso de IA generativa

Os autores declaram que não utilizaram ferramentas de inteligência artificial generativa na elaboração deste trabalho.

Referências

Angermann, H., & Ramzan, N. (2017). Taxonomy matching using background knowledge: Linked data, semantic web and heterogeneous repositories. Springer International Publishing.

Bailey, D. (2004). An efficient Euclidean distance transform. In Proceedings of the 11th International Semantic Web Conference (pp. 394–408). Springer.

Barbosa, N. T. (2021). Para uma economia da informação semântica: A construção de ambientes semânticos para a recuperação inteligente da informação (Tese de Doutorado, Universidade Federal Fluminense).

Barbosa, N. T., & Campos, M. L. de A. (2022). Diretrizes para a compatibilização de SOCs com vistas a uma recuperação inteligente da informação. Scire: Representación y Organización del Conocimiento, 28, 67–81.

Cheatham, M., & Hitzler, P. (2013). String similarity metrics for ontology alignment. In Proceedings of the International Semantic Web Conference (ISWC 2013) (pp. 294–309). Springer.

Dahlberg, I. (1981). Towards establishment of compatibility between indexing languages. International Classification, 8(2), 88–91.

Dahlberg, I. (1983). Terminological definitions: Characteristics and demands. In D. Duquet-Picard (Ed.), Problèmes de la définition et de la synonymie en terminologie:

Actes du colloque international de terminologie, Université Laval, Québec, 23–27 mai 1982 (pp. 13–51). Association Internationale de Terminologie.

Feng, Y., Bagheri, E., Ensan, F., & Jovanovic, J. (2017). The state of the art in semantic relatedness: A framework for comparison. Knowledge Engineering Review, 32, 1–30. https://doi.org/10.1017/S0269888917000029

Hamming, R. W. (1950). Error detecting and error correcting codes. Bell System Technical Journal, 29(2), 147–160. https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. Morgan & Claypool. https://doi.org/10.2200/S00639ED1V01Y201504HLT027

Hellinger, E. (2009). New formation of the theory of square forms of infinite number of Veraenderlichen. Journal für die reine und angewandte Mathematik, 210–217.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11–21. https://doi.org/10.1108/eb026526

Kernighan, B., & Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. Bell System Technical Journal, 49, 291–307.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady, 10, 707–710.

Lévy, P. (2014). A esfera semântica: Tomo 1 – Computação, cognição, economia da informação. Annablume.

Lévy, P. (2016). Blog. https://pierrelevyblog.com/

Lévy, P. (2019). A metalinguagem da Economia da Informação – white paper. Pré-print. Documento técnico não publicado. DOI: 10.13140/RG.2.2.11232.33281.

Lévy, P. (2023). Semantic computing with IEML. Collective Intelligence Volume 2:4: 1–28. DOI: 10.1177/26339137231207634.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Addison-Wesley.

Tanenbaum, A. S. (2003). Redes de computadores (4ª ed.). Elsevier.

Wei, Qiyao, Morrell, Edward, Goetz, Lea, & Schaar, Mihaela van der. (2025). Semantic-KG: Using Knowledge Graphs to Construct Benchmarks for Measuring Semantic Similarity. 39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Wu, Z., & Palmer, M. (1994). Verbs semantics and lexical selection. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (pp. 133–138). ACM