Diretrizes para a compatibilização de SOCs com vistas a uma recuperação inteligente da informação

Directrices para la compatibilidad del SOC con miras a la recuperación inteligente de la información

Guidelines for KOS compatibility towards intelligent information retrieval

Nilson Theobald BARBOSA (1), Maria Luiza de Almeida CAMPOS (2)

(1) Universidade Federal do Rio de Janeiro - Cidade Universitária, Rio de Janeiro/RJ, nilson@tbarbosa.org
(2) Universidade Federal Fluminense – Niterói/RJ, Universidade Federal da Bahia – Salvador/BA, marialuizalmeida@gmail.com

Resumen

Se presenta un enfoque para la compatibilización de vocabularios heterogéneos con el fin de permitir la recuperación inteligente de información en diferentes bases de datos, asegurando que los vocabularios originales se mantengan sin cambios. Este estudio se caracteriza por un enfoque cualitativo que supone un desarrollo interpretativo de los datos recogidos a partir de la investigación bibliográfica y documental. Como producto de esta investigación se presenta un conjunto de directrices que, apoyadas en métodos, técnicas y algoritmos computacionales, apuntan a la posibilidad de crear procesos automatizados de compatibilidad semántica de vocabularios que conduzcan a un proceso inteligente de recuperación de información distribuida en diferentes SOCs.

Palabras clave: Sistemas de organización del conocimiento. Interoperabilidad semántica. Lenguaje intermedio. Coordenadas semánticas. Recuperación de información.

Abstract

An approach to the compatibilization of heterogeneous vocabularies is presented aiming to allow the intelligent retrieval of information in different databases, ensuring that the original vocabularies are kept without change. This study is characterized by a qualitative approach that supposes an interpretative development of data collected from bibliographic and documentary research. As a product of this research, a set of guidelines is presented that, supported by methods, techniques, and computational algorithms, points to the possibility of creating automated processes of semantic compatibilization of vocabularies that may lead to an intelligent process of retrieval of information distributed in different KOS.

Keywords: Knowledge organization systems. Semantic interoperability. Intermediate language. Semantic coordinates. Information retrieval.

1. Introdução

Já há algum tempo vivemos no mundo dos dados digitais, com a evolução das tecnologias criadas há aproximadamente meio século que dariam início a mais uma revolução em nossa história e permitiram o surgimento da Internet. Nos tempos atuais parece que nossa capacidade de criar e produzir dados ultrapassa de longe nossa capacidade de gerenciar e permitir que estes dados, além de estarem acessíveis, sejam compreendidos, enfim que façam sentido e sejam fonte de informações para quem delas necessita. Gantz e Reinsel (2010) mostravam em um relatório que entre 2010 e 2020 o total de registros digitais criados e replicados pelo mundo teriam uma evolução para inconcebíveis 35 trilhões de gigabytes. Esta previsão foi confirmada em 2020 e temos hoje uma perspectiva de atingirmos perto de 150 zettabytes em 2024, considerando o volume de dados criados e capturados em todo o mundo (Statista, 2021).

Todo o avanço tecnológico e computacional que contribui com a geração deste volume de dados e cria redes com maior desempenho, sejam as redes de fibra ótica, as redes sem fio ou a vindoura rede 5G, e que usa repositórios “infinitos” e fornece computação de alto poder de desempenho na palma da mão tem, apesar disso, um limitador de forte impacto para a utilização plena de toda esta possível informação. Uma miríade de repositórios de dados e seus conteúdos com todos os tipos de dados se multiplicam exponencialmente. Estes repositórios, seus dados e suas linguagens de indexação heterogêneas em todos os seus níveis, seja dentro das organizações seja na Web como um todo, ainda não são capazes de oferecer aos seus usuários uma recuperação plena e semântica da informação que está disponível.

Para que possamos ter pleno usufruto deste capital de conhecimento, um forte limitador nos persegue, aliás desde antes da criação da Internet, que é a incapacidade da humanidade de resolver seus problemas de divisão culturais e linguísticos

e fundamentalmente de resolver a incompatibilidade de seus sistemas de classificação. Pierre Lévy chama este processo envolvendo um enorme crescimento computacional, difusão de dados, produção de registros e consumo crescente de informações digitais de 'memória digital participativa', afirmando que esta memória está apenas em vias de constituição, a despeito de todo avanço tecnológico. Temos como desafio a automatização das operações cognitivas de análise e interconexão das informações que supostamente estão disponíveis. Não sabemos, ainda, como transformar de forma sistemática e efetiva este oceano de dados em conhecimento e menos ainda como transformar o meio digital em observatório reflexivo de nossas inteligências coletivas (Lévy, 2014).

A interoperabilidade física entre computadores e entre bases de dados é uma questão já bem resolvida pela tecnologia, a compatibilização entre termos de mesma grafia em diferentes vocabulários já é bem realizada pelo alto desempenho computacional disponível e facilitada pela conectividade física, mas transpor a barreira semântica, em que precisamos compatibilizar conceitos, tenham eles a mesma expressão verbal ou não, é uma questão ainda a ser resolvida.

Apesar de parecer tentador criar sistemas de indexação e vocabulários unificados para resolver este problema, em um ambiente aberto e não controlado nem sempre é possível recorrer a esta solução, e precisamos partir para soluções que possibilitem recuperar informações de bases indexadas por sistemas heterogêneos sem fazer alterações nestas bases ou em seus vocabulários através do estabelecimento de correspondências e mapeamentos de conceitos, e não simplesmente de termos verbais, entre estes vocabulários.

Portanto, esta é, em síntese, a questão que perseguimos aqui. A possibilidade de compatibilização semântica automatizada de diferentes sistemas de organização do conhecimento sem a alteração de seus ambientes originais, com a utilização de metalinguagens, que parecem ser capazes de fornecer as bases teóricas para o desempenho desta tarefa. Estas metalinguagens devem ser formadas como uma linguagem intermediária entre os diferentes vocabulários fonte possibilitando diferentes atores navegarem por esta linguagem e, de forma contextual e semântica, recuperarem a informação pretendida, não mais com base em simples comparações de cadeias de caracteres, mas sim em seu significado.

A seguir colocaremos o problema brevemente relacionado à criação de um ambiente para uma recuperação inteligente da informação, para depois apresentarmos as diretrizes que consideramos poder auxiliar na elaboração deste ambiente. Apresentaremos em sequência uma síntese das técnicas que podem ser utilizadas para o estabelecimento de equivalências e proximidades semânticas entre os conceitos e por fim as considerações finais.

2. Um espaço para a recuperação inteligente da informação: as linguagens intermediárias e as coordenadas semânticas

Após a disseminação da computação, da Internet e da Web, estamos diante de muitas iniciativas que procuram resolver a disparidade e heterogeneidade entre os sistemas de informação. Soergel (1972, 1974) já abordava esta questão e apresentava discussão que endereça estes problemas de compatibilidade. Uma solução apresentada, para enfrentar a abundância de tesauros e a contínua criação de novos, seria a criação de um Tesauro Fonte Universal, armazenado em computador, onde os elementos de todos os tesauros existentes poderiam ser coletados, assim como a indicação de todas as suas relações. Apesar de este tesauro universal poder ser usado como uma fonte de informação para descritores existentes e relações entre seus conceitos, e para criação de novos sistemas utilizando os conceitos existentes, estabelecer este tesauro universal seria um grande empreendimento e sua realização poderia apenas criar mecanismos de compatibilidade únicos para os sistemas que originem seus elementos da fonte comum (Dahlberg, 1981).

Como um caminho de compreensão e solução do problema, nosso olhar se volta inicialmente para a norma internacional que trata de interoperabilidade entre vocabulários e que oferece linhas de atuação para obter este fim. Observamos que a norma ISO 25964 parte 2 elenca diferentes modelos estruturais para a realização de mapeamentos entre vocabulários, a saber, Unidade estrutural, Ligação direta e Estrutura central, e estes modelos apresentados reforçam a visão do caminho para um vocabulário único, uma vez que no primeiro caso não chegamos sequer a ter mapeamentos, no segundo caso temos a proposta dos mapeamentos um-a-um, unidirecionais, custosos e difíceis de implantar, e no terceiro caso temos a utilização ou criação de um vocabulário central, que além de também ser usado para indexação é usado como um possível comutador entre outros vocabulários menores ou mais específicos. Portanto, se tomarmos por base o documento padrão que se propõe a normatizar os processos de interoperabilidade entre sistemas de organização do conhecimento somos levados a estabelecer processos de compatibilização que levam à criação de vocabulários únicos.

Trilhando um caminho diferente, nosso interesse se volta para uma abordagem de criação de linguagens intermediárias como uma estratégia para a compatibilização de vocabulários, por considerar que as necessidades de compatibilização de linguagens para o momento atual, com vistas a uma Web Semântica devem ter por base este arcabouço teórico da Ciência da Informação.

As discussões que embasam a proposta de métodos de léxicos intermediários para a compatibilização de vocabulários remontam aos trabalhos publicados por Hammond e Rosenborg (1962), Newman (1965) e Henderson et al.(1966), tendo a questão da compatibilidade e conversibilidade recebido especial atenção no relatório da UNESCO de 1971. Neste relatório temos uma definição de compatibilidade como sendo "uma qualidade de sistemas cujos produtos podem ser intercambiados, apesar de suas diferenças de notação, estrutura, suportes físicos etc., sem qualquer mecanismo especial de conversão" (Unesco, 1971). Além disso, conversão é definida como "o processo de transformar registros de informação, com respeito à codificação, estrutura de dados etc., de modo a fazê-los intercambiáveis entre dois ou mais sistemas usando diferentes convenções" (Unesco, 1971, p.147).

Uma importante contribuição foi também desenvolvida por J. C. Gardin (1967, 1973) e pelo seu grupo de trabalho na França, definindo que um léxico intermediário é destinado a acessar documentos indexados em termos de uma linguagem de indexação para outra sem que haja a perda de informação. Ele implica no mapeamento de dois ou mais vocabulários para uma linguagem intermediária ou neutra.

As investigações empíricas apresentadas por Wellisch (1972), Agraev et al.(1974), Smith (1974), Svenonius (1975) e Wersig (1975) apresentam estudos que definem a natureza das linguagens de indexação comparadas, a metodologia para comparação de linguagens de indexação e a estrutura dos elementos das linguagens de indexação mais adequadas para intercambiamento, sendo de especial interesse os estudos de Horsnell (1975) sobre a criação de um "léxico intermediário" (Dahlberg, 1981).

Considerando as abordagens seminais estudadas, a definição de linguagem intermediária vista e aceita aqui será como mostrada por Dahlberg (1981) e Neville (1970, 1972), baseada em uma codificação de conceitos, que permite o estabelecimento de uma equivalência conceitual de descritores de diferentes linguagens (Campos et al., 2009) e, de forma compatível com esta definição, a premissa aqui assumida é a importância e a necessidade de se realizar qualquer processo de compatibilização, alinhamento e mapeamento de vocabulários sem que os vocabulários originais sejam alterados ou tenham suas características modificadas, considerando as grandes dificuldades, especialmente administrativas e políticas, de conseguir realizar estas tarefas.

Um dos modos de realizar a construção deste dispositivo é a utilização do método da matriz de compatibilização conceitual de Dahlberg (1981). Partindo de seu método analítico-sintético, Dahlberg propõe a construção de uma matriz representativa da compatibilidade conceitual entre sistemas ordenados. Esta matriz é um mapeamento da potencialidade semântica das linguagens a serem compatibilizadas e, a partir daí, pode fornecer os resultados da análise de compatibilidade entre estas linguagens sob os pontos de vista sintático, estrutural e semântico.

Nesse sentido, também o método de Neville (1972) chamado de reconciliação de tesauros, tem por base o mesmo princípio de construção de léxicos intermediários apresentados por Natacha Gardin (1969) e Coates (1970), pressupondo que a compatibilização dos sistemas de organização do conhecimento deve considerar não só a sintaxe dos termos descritores, mas também os seus conteúdos conceituais, isto é, suas significações, expressas pelas suas definições (compatibilidade semântica). Este método prevê a elaboração de uma linguagem intermediária, baseada na codificação numérica de conceitos (onde cada conceito poderia ser identificado por um código numérico), possibilitando (i) estabelecer equivalência conceitual entre termos descritores de diferentes linguagens e, (ii) realizar a conversão automática de termos equivalentes e de termos específicos para genéricos (Bocatto e Torquetti, 2012).

Também nesta direção, consideramos de grande interesse para nosso trabalho discutir o conceito de "Sistema de Coordenadas Semânticas", apresentado por Pierre Lévy (Lévy, 2014, p. 312, 2019, p. 27), cuja formulação e procedimentos metodológicos permitem uma aproximação com o conceito de dispositivos de comutação, e consequentemente com os léxicos intermediários, colocando uma visão atual e centrada em procedimentos computacionais e automáticos para esta implementação.

Pierre Lévy vem se dedicando a explicitar uma construção teórica, que ele denomina de IEML – Metalinguagem da Economia da Informação, e 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. Ao propor esta construção, ele apresenta a IEML principalmente como uma

linguagem artificial que se traduz automaticamente em línguas naturais (Lévy, 2014).

Neste sentido, na construção teórica proposta por Lévy, nos interessa sobremaneira o conceito de linguagem ponte e de sistema de coordenadas semânticas apresentadas na IEML. Como linguagem ponte podemos entender uma linguagem artificial ou natural usada como uma linguagem intermediária para tradução entre muitas línguas diferentes – para traduzir entre qualquer par de idiomas A e B, uma função traduz A para a linguagem ponte L, depois de L para B. Como sistema de coordenadas semânticas é importante entender que sua função seria de permitir um sistema de endereçamento que possibilite computar as relações e distanciamentos semânticos existentes entre as linguagens (Lévy, 2014).

Entendemos que a totalidade da proposta de Lévy apresenta proposições de difícil implantação dada a sua complexidade e extensão, mas por outro lado, apresenta caminhos metodológicos para a construção de dispositivos que permitam comutação entre diferentes ontologias que particularmente interessam em nossa pesquisa. Uma destas propostas é o estabelecimento de identificadores únicos, chamados de Uniform Semantic Locators (USL). As operações calculáveis realizadas nos conjuntos de sequências que são os USL são ao mesmo tempo operações realizadas sobre o sentido que estes conjuntos representam (ou seja, os conceitos). “A principal ideia a ser retida é a de que um caminho qualquer no espaço hipertextual das conexões entre USL pode ser representado por uma função e a de que essa função pode ter uma pertinência semântica” (Lévy, 2014, p.478).

Dessa forma, conforme Lévy, a esfera semântica e a linguagem IEML funcionam como um sistema de codificação do sentido concebido para tornar automaticamente calculáveis operações sobre os conceitos e sobre suas operações semânticas, e tudo isso repousa, na prática, sobre a existência de um conjunto de circuitos semânticos matriciais funcionando como convenção de tradução dos textos IEML para os circuitos semânticos selecionados em línguas naturais e vice-versa.

Compreendemos que a proposta de Lévy é uma criação desta convenção de tradução que avance para todas as coisas existentes, representando um sistema de comutação universal, mas defendemos também que esta proposta pode ser analisada sob o ponto de vista da criação de dispositivos cujo funcionamento é coerente com as propostas de construção de linguagens intermediárias, ou seja, dispositivos de comutação, apresentadas pela Ciência da Informação para a recuperação da informação em ambientes heterogêneos.

Para isso recuperamos a definição de que a unidade básica da IEML, o USL, não se limita a descrever um conceito, mas pode ser usada para intermediar consultas em uma base dados. Neste ambiente com uma coleção de identificadores únicos (ou USL), é possível calcular os mais semelhantes a outros USL, ou seja, os mais representativos da coleção, e aqueles que têm menos em comum com outros membros da coleção, ou seja, a maioria dos dispositivos da coleção.

A proposta de Lévy nos leva a compreender que os conceitos e suas representações dentro de um sistema de comutação para recuperação da informação entre várias linguagens podem ser recuperados através dos métodos de mapeamento que sejam voltados para gerar linguagens intermediárias e podemos ver que em sua proposta é necessário que estes identificadores únicos apresentem descritores que representem seu significado semântico, o que Dahlberg em seu trabalho procurou descrever através do Registro do Conceito (Dahlberg, 1981).

Portanto, os processos computacionais de alinhamento e mapeamento que são capazes de compatibilizar SOCs e seus termos podem ser utilizados para a criação de uma “esfera semântica”, não universal e global, mas que atenda à recuperação da informação em ambientes com diversos sistemas de indexação. A definição dos “USL”, baseadas no Registro de Conceito de Dahlberg, para estes ambientes, pode levar a um processo de recuperação da informação que seja capaz de fornecer a um usuário interessado em utilizar este ambiente de multivocabulários o significado de cada conceito, sua representação em cada linguagem utilizada, e a medida de distância semântica de cada conceito para outros conceitos, iguais ou semelhantes. É este objetivo que perseguimos na formulação das diretrizes que apresentamos a seguir.

3. Diretrizes para elaboração de Linguagens Intermediárias entre SOCs

No contexto deste artigo, apoiados na proposta de Nurcan et al.(1999), vamos destacar alguns aspectos na apresentação das Diretrizes (DIR01 a DIR06), ou seja, iremos denominá-las como um objetivo a ser atingido, logo após apresentaremos uma descrição apontando a sua finalidade e importância e por último apresentaremos instruções visando alcançar os propósitos esperados enunciados na denominação da Diretriz.

3.1. DIR 01: Desenvolvimento e manutenção dos sistemas de organização do conhecimento por profissionais especializados

Num momento em que, tanto os novos vocabulários criados, como a imensa quantidade de sistemas já existentes, precisam participar de processos automatizados de compatibilização de informações com vistas à sua recuperação entre sistemas heterogêneos, a questão da utilização de padrões sólidos é questão essencial para que este fim seja alcançado.

Nesse sentido, a construção destes sistemas de organização do conhecimento, talvez hoje mais do que nunca precisem ser desenvolvidos por profissionais altamente qualificados para sua construção, pois, caso contrário, todos os esforços na criação de técnicas modernas, algoritmos de alta eficiência e computadores e redes de alto desempenho, não serão capazes de avançar na tarefa de interoperar sistema heterogêneos. De fato, o desenvolvimento de Sistemas de Organização de Conhecimento deve ser um trabalho conjunto entre aqueles que dominam os processos de classificação de domínio, os especialistas do domínio, e profissionais de TI.

É importante notar que esta recomendação não se aplica apenas no momento da criação dos sistemas, mas deve ser seguida por toda sua vida, em seus processos de manutenção e atualização, e da mesma forma se aplica tanto a taxonomias e tesauros, quanto a construção de ontologias, pois o conhecimento referente ao tratamento e organização da informação é essencial para ser combinado com a utilização das novas tecnologias e novas linguagens de representação de sistemas de organização do conhecimento.

3.2. DIR 02: Utilização de linguagens de representação padrão, compatíveis e abertas para construção dos sistemas de organização do conhecimento

Proporcionar aos sistemas de organização do conhecimento participantes capacidade de serem lidos e interpretados por agentes de software de forma compatível com representações padrão definidas pelas tecnologias da web semântica, em especial aquelas defendidas pelo consórcio W3C.

As tecnologias ligadas à web semântica apresentam múltiplas opções para representação de dados e temos diversas possibilidades para representação de sistemas de organização do conhecimento dentre RDF, RDF-S, OWL, SKOS, entre outras.

A análise do modelo SKOS nos mostra que este formato, pela sua difusão e pela facilidade de conversão entre diferentes modelos, inclusive a partir de vocabulários e tesauros que possuem apenas um modelo de dados descritivo, nos permite dizer que este pode ser considerado um modelo preferencial para a representação de sistemas de organização do conhecimento em geral, e em especial, tesauros e taxonomias.

Outro aspecto importante é a recomendação por este modelo de representação assumido pelo World Wide Web Consortium, sob a justificativa que sua difusão mundial e a sua representação em RDF proporciona um padrão facilmente interoperável entre diferentes instituições.

No caso de SOCs ainda não representados em SKOS é uma boa prática realizar sua conversão (utilizaremos aqui como exemplo um tesauro por oferecer mais elementos para consideração, mas o procedimento se aplica a diferentes tipos de SOC) para o modelo SKOS. Ao transferir toda sua base de conhecimento e suas relações estruturais podemos utilizar alguns procedimentos, baseados nas recomendações e padronizações do W3C, que objetivam converter o SOC em questão para uma codificação SKOS/RDF.

Estes procedimentos se iniciam com a análise do tesauro em questão de forma a verificar as suas relações padronizadas (tais como, TG, TGP, TE, TEP e TA) e as não-padronizadas. Como resultado deste passo teremos um catálogo de todos os itens de dados e todas as restrições, como uma lista de todas as características do tesauro.

Em seguida, todas estas características devem ser mapeadas para o formato SKOS RDF. Nesse momento se define como cada item de dados será representado no esquema SKOS, gerando uma tabela de cada item do tesauro para o item SKOS que o represente.

Por fim, um especialista em tecnologia da informação deve ser capaz de realizar a função de elaboração de um programa de conversão que gere o SOC em seu formato SKOS RDF. Nesse processo não há interferência nos dados originais e os dois vocabulários têm as mesmas informações, apenas representados de forma diferente, mas facilitando sobremaneira a utilização de agentes para compatibilização, pela sua representação em uma linguagem de dados padronizada, voltada para ser usada por programas de computador.

Como resultado deste processo teremos SOCs capazes de participar de um processo de compatibilização, mesmo organizados em diferentes instituições ou departamentos, permitindo a ação de agentes de software que realizem sua leitura e criação dos objetos registros de conceito.

3.3. DIR 03: Utilização de definições para os conceitos nos SOC

Possibilitar um ganho na identificação do significado semântico de cada conceito a ser compatibilizado, a partir da explicitação de suas características ou propriedades, sob a forma de definição conceitual, que possa ser extraída e tratada por agentes de software.

As notas de escopo/aplicação, onde as definições de um conceito são apresentadas em tesauros, são úteis para um processo de recuperação ou de compatibilização manual. Ao se propor uma situação de compatibilização semântica por processos automáticos elas ganham uma nova importância, uma vez que a utilização de técnicas e tecnologias baseadas em algoritmos computacionais podem ser aplicados e estas informações serem comparadas entre diferentes conceitos.

Diversas técnicas podem ser usadas para a comparação dos textos livres e não estruturados usados para compor as definições (que se encontram nas notas de escopo/aplicação), que por sua vez podem estabelecer parâmetros para determinar a aproximação semântica e a similaridade entre dois conceitos, adicionalmente às informações estruturais extraídas do SOC. Uma destas técnicas é a análise de distribuição semântica, que determina esta similaridade como resultado da similaridade da distribuição linguística (Boleda, 2020). Outra técnica atual é a de word embedding, que agrupa na verdade um conjunto de técnicas para mapear de forma sintática e semântica um texto em linguagem natural, com a utilização de meios estatísticos. Como resultado, palavras de um texto são levadas para um espaço vetorial e podem ser comparadas com palavras de mesmo conteúdo semântico em outro texto, a partir da criação de um embedding space, que represente semanticamente as palavras determinantes do sentido de cada um dos textos (Goldberg, 2017).

O emprego de tais técnicas nas notas de escopo vai no sentido de estabelecer uma representação semântica destas notas, através da extração de frases e palavras que representem seu significado e permitir que esta informação seja utilizada em conjunto com a expressão verbal originalmente comparada, com as comparações que utilizam a estrutura e taxonomia do SOC e com as comparações que utilizam os termos associados. A utilização das técnicas que traduzem a interpretação semântica das notas de escopo pode elevar sua utilização de simples texto de apoio para usuários humanos dos SOC para importantes meios de estabelecer conexão semântica entre conceitos.

Desta forma, nosso objetivo aqui é determinar que a inclusão de notas de escopo para o maior número possível de conceitos nos vocabulários representa esforço decisivo para que o processo de automação da compatibilização e correspondência destes conceitos ocorra de forma mais eficiente e precisa possível. Assim, as informações geradas da nota de escopo através das técnicas de distribuição semântica e word embedding deveriam ser incluídas ao registro de conceito (ver diretriz DIR 05 a seguir) para cada conceito analisado e assim este parâmetro ser também utilizado para a definição da similaridade e distância semântica entre os conceitos.

Além disso, normalmente este campo é utilizado de forma livre e não estruturada, de forma coerente com o que é apontado pela norma ISO 25964, mas a inclusão na norma de uma proposta de futura padronização deste campo, transformando-o em um campo estruturado com informações que pudessem, de forma imediata e automática, participarem do processo de compatibilização, traria ganho significativo para a compatibilização dos SOC mais comumente utilizados, tais como os tesauros.

Assim, um atributo comum, como as notas de escopo, já largamente utilizado para o aumento das capacidades semânticas dos conceitos de um SOC pode ser trazido para compor o sistema proposto aqui, de coordenadas entre vocabulários, aumentando as possibilidades de mapeamentos com alto grau de acerto e acurácia.

3.4. DIR 04: Estabelecimento de identificadores únicos para os registros de conceito

Garantir que a representação dos conceitos, em especial os registros de conceito, que forem gerados de forma automática possam ser identificados de forma única, mesmo em um ambiente aberto na Internet.

Para Dahlberg (1981), um dos requisitos para a criação do registro do conceito, abordado anteriormente, é a utilização da notação utilizada pelo SOC para representar este conceito e possibilitar que participe do processo de compatibilização através da criação de suas matrizes de compatibilidade semântica, sendo de grande importância e um requisito ao processo de compatibilização.

Se num ambiente manual esta identificação notacional dos conceitos já era considerada importante, num ambiente de criação de um espaço semântico gerado de forma automática por agentes de software, passa a ser essencial e condição básica para sua implementação.

Ao abordar e propor a criação da esfera semântica e o espaço de coordenadas semânticas que a compõe, Pierre Levy (2014) também ressalta a

necessidade imperativa de estabelecer um sistema de identificação e endereçamento dos conceitos. Para isto, o autor justifica esta posição afirmando que, com relação ao meio digital, a única certeza que temos é que sua história acaba de começar, ao estabelecer um vetor de crescimento do processo de codificação digital que vivemos em nossa história recente.

Neste sentido, 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, Levy, propõe a criação dos Uniform Semantic Locators (USL), cuja função é a identificação e endereçamento dos conceitos, com a utilização da Information Economy Meta Language (IEML) (Lévy, 2014).

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 URLs. 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 suas URL.

A formação desta esfera semântica completa e unifica, segundo Levy, a ação dos autômatos processadores de símbolos, por sua vez interconectados pela internet, com o conjunto dos dados interconectados pela web. A introdução desta nova camada de endereçamento possibilita a interconexão dos dados, criando uma forma de sinergia diferente daquela que se conhece hoje. O sistema de endereçamento virtual proposto pela IEML define que cada USL distinto codifica um conceito distinto. Como cada conceito pode ser traduzido em línguas naturais, essa identificação na metalinguagem funciona como uma linguagem pivô entre as línguas e os sistemas simbólicos naturais.

Definimos aqui a tomada deste caminho, onde cada conceito precisa ter sua codificação única, mas é necessário fazer sua representação sem que tenhamos ainda disponível um sistema de codificação global, como proposto por Levy na IEML, mas ainda assim estabelecer meios para que os conceitos que formarão nosso espaço semântico possam ser representados por identificadores únicos.

Desta forma, como não temos disponível uma codificação global, e os simples URL que ligam páginas na web, não servem para nosso propósito, recorremos à utilização de um Uniform Resource Identifier (URI), que precisa se apoiar em uma construção do tipo URL, mas pode ser capaz de estabelecer códigos de identificação únicos para objetos na Web.

Para a formação do URI de cada conceito, contendo suas informações, podemos, portanto, partir das notações já definidas em cada SOC, ou estabelecer uma notação sequencial caso um determinado sistema não a possua. Desta forma, o SRI responsável por realizar as buscas descentralizadas pode ser capaz de funcionar utilizando todo o espaço semântico disponível (ver diretriz DIR 05, a seguir) de forma inteligente.

Desta forma, os URI dos registros de conceito gerados deverão ser constituídas minimamente pelos seguintes campos: a) protocolo utilizado (http ou https) e domínio da instituição ou setor detentor e responsável pelo vocabulário de origem; b) nome do vocabulário; e c) identificador único extraído do vocabulário, ou numeração sequencial gerada durante o processo de geração dos registros de conceito.

Como resultado deste processo temos a capacidade de estabelecer, utilizando as camadas disponíveis do endereçamento digital, um identificador único para os conceitos que farão parte de nosso espaço semântico, e que permitirá suas interligações semânticas propostas a seguir.

3.5. DIR 05: Elaboração do espaço semântico a partir da extração dos registros de conceito e seu mapeamento semântico

Elaborar um espaço semântico a partir da extração das informações que farão parte da identidade de cada conceito, da criação dos registros de conceito, e da realização de um mapeamento semântico entre os conceitos de diferentes SOC, que possam vir a ser utilizados em um sistema de recuperação inteligente.

Esta diretriz foi construída a partir dos passos seguidos pelos experimentos de compatibilização realizados em Barbosa (2021), onde pudemos reproduzir caminhos a serem seguidos por um agente de software ao realizar um processo de compatibilização e correspondência.

Para a composição dos registros de conceito, de forma a atender aos processos de compatibilização, propomos a inclusão básica dos seguintes campos, que serão extraídos dos sistemas de organização do conhecimento: (i) expressão verbal do conceito; (ii) notação do conceito extraída do sistema de organização do conhecimento ou gerada de forma automática e sequencial; (iii) termo genérico maior, ou seja, o nível mais abrangente de sua escala hierárquica; (iv) termo genérico imediato; (v) termos específicos; (vi) termos associados; e (vii) definição extraída de sua nota de escopo.

Os passos propostos são especialmente voltados para um processo cujo objetivo não é simplesmente estabelecer um apontamento, ou um mapeamento de um determinado conceito para outro numa situação um para um, e de mesma forma não tem como objetivo a fusão, junção ou integração de vocabulários, mas sim criar objetos denominados registros de conceito que, de forma digital, armazenem as informações semânticas para um determinado conceito em um determinado SOC e, além disso, estabeleçam relações de apontamento para conceitos de outros SOC que possam ser relacionados semanticamente, visando possibilitar interoperação em espaços semânticos. Conforme mostrado em Barbosa (2021), estas relações de apontamento e mapeamento propostas aqui não se limitam a identificar equivalências em conceitos com expressões verbais iguais ou similares, mas identificar equivalência conceitual mesmo com diferentes formas verbais e, de mesma forma, serem capazes de identificar que determinados conceitos com mesma expressão verbal podem ter significados semânticos completamente diferentes e por isso não podem ser mutuamente mapeados.

O processo que recomendamos aqui tem por objetivo a utilização das técnicas e algoritmos computacionais descritos em Barbosa (2021, p. 169), para que seja possível o projeto de agentes de software que não simplesmente estabeleçam ponteiros interligando conceitos similares semanticamente, mas que sejam capazes de determinar quanto um determinado conceito se relaciona a outro, ou seja, se apresenta total compatibilidade semântica, ou uma compatibilidade parcial, determinada num intervalo entre 0 e 1, ou seja,

0 < similaridade semântica calculada <= 1

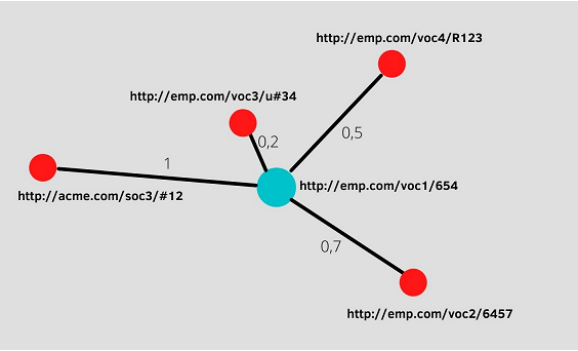

Figura 1. Correspondências semânticas a partir de um conceito (Barbosa, 2021)

Na figura 1 podemos ver que os conceitos foram identificados através de uma URI e, por exemplo, o conceito http://emp.com/voc1/654 foi mapeado para quatro outros conceitos de diferentes vocabulários, e suas medidas de similaridade semânticas para cada um deles foram explicitadas pelas suas relações de ligação. O espaço semântico a ser criado é um espaço multidimensional onde todos os conceitos extraídos se relacionam por estas distâncias semânticas.

Para a operacionalização deste processo, de acordo com as propostas de Neville (1972) e Dahlberg (1981), o primeiro passo é buscar pela identidade sintática entre as expressões verbais presentes nos vocabulários, assumindo como pressuposto sua possível identidade semântica. Como mostramos em nosso experimento e nas discussões teóricas sobre as possíveis e diversas técnicas usadas (Barbosa, 2021), esta identidade sintática se inicia pela total igualdade de caracteres, passando pela pesquisa de subcadeias de caracteres, plurais, formas verbais, e assemelhados.

Para isso, o primeiro passo a ser tomado, para cada expressão verbal de cada vocabulário a ser compatibilizado, é extrair de sua estrutura as informações que irão compor seu registro de conceito, tais como, a própria expressão verbal, sua notação no vocabulário, o termo genérico maior, o termo genérico, os termos específicos, os termos associados e sua nota de escopo.

Após este procedimento inicial, as operações se voltam para dois procedimentos básicos, ou seja,

1. verificar se as formas verbais que se equivalem sintaticamente representam conceitos que são semanticamente iguais ou semelhantes, ou se trata de polissemias, e
2. descobrir nos vocabulários participantes possíveis expressões verbais que, apesar de dessemelhantes sintaticamente, apresentam similaridade semântica que as tornem capazes de serem mapeadas dentro do espaço semântico construído.

Portanto, para chegar a estes propósitos, os passos a serem seguidos, relacionando e detalhando as ações a serem implementadas, são:

1. para cada um dos registros gerados, identificar registros de conceito nos outros vocabulários participantes que possuam a mesma expressão verbal;
2. 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, listadas com mais detalhes na seção 4 deste trabalho.

3. 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 (2), em seus termos genéricos, termos específicos e termos associados;
4. 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 - observar aqui os procedimentos realizados para os mapeamentos descritos em Barbosa (2021);
5. 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;
6. 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;
7. armazenar as distâncias semânticas calculadas para cada conceito apontado, nos próprios registros de conceito, utilizando os identificadores únicos para representar os conceitos mapeados (figura 1).

Na figura 2 mostramos os sistemas de organização do conhecimento S1, S2, ..., Sn, que respectivamente são utilizados para indexar as bases B1, B2, ..., Bn, e que são percorridos pelo agente de software AS, responsável por executar as operações detalhadas acima e criar o espaço semântico ES.

O que apresentamos aqui, portanto, como caminho a ser seguido não é o estabelecimento de uma matriz de mapeamento booleano, onde a ligação entre conceitos existe ou não existe. O caminho apresentado é o estabelecimento de uma base de dados de registros de conceitos, representando um grupo de SOCs participantes que, ao ser gerada a partir dos caminhos propostos aqui vai ser utilizada por um sistema de recuperação inteligente da informação (diretriz 06), oferecendo uma visão semântica do conjunto de sistemas de organização do conhecimento compatibilizados e servindo de base para buscas interativas inteligentes entre os diversos SOCs por parte dos usuários do sistema, como veremos a seguir.

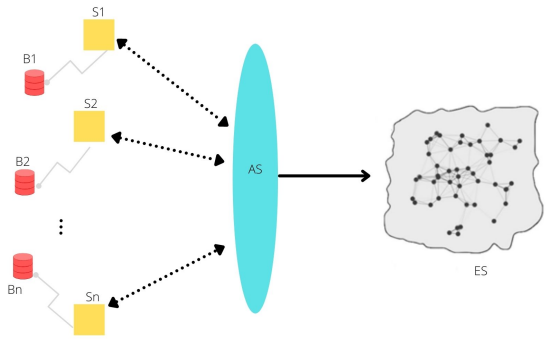

Figura 2. Sistema de criação do espaço semântico (Barbosa, 2021)

3.6. DIR 06: Estabelecimento de um espaço de recuperação inteligente da informação que trabalhe com os registros de conceitos e suas distâncias semânticas

Estabelecer um sistema de recuperação da informação que faça uso do espaço semântico gerado onde conceitos interconectados pelas similaridades semânticas calculadas possam oferecer buscas semânticas inteligentes em ambientes heterogêneos multivocabulários.

Para que seja possível fazer uso efetivo dos processos de compatibilização, correspondência e mapeamento de vocabulários e seus conceitos, estes processos devem ser voltados, como já defendemos anteriormente neste trabalho, para que seja possível a realização da recuperação inteligente da informação, onde se procura suplantar as barreiras impostas pela diversidade e heterogeneidade dos sistemas de organização da informação utilizados para indexar documentos e fornecer informações relevantes para os usuários.

Para isto, nossa proposta se cristaliza na construção de ambientes de recuperação da informação que façam uso do espaço semântico proposto anteriormente, resolvendo a heterogeneidade, atingindo uma interoperabilidade semântica entre vocabulários e adicionalmente trazendo para as mãos do usuário a possibilidade de interagir e definir os limites de compatibilidade que interessam aos seus propósitos.

Este sistema de recuperação da informação ao ser acionado por um usuário ao percorrer a hierarquia de um vocabulário, ou mesmo a partir de um termo livre, poderá ser capaz de identificar os conceitos que atendem àquela busca em outros vocabulários que participem do mesmo espaço semântico, oferecendo conceitos similares nos

outros vocabulários e sendo capaz de afirmar para o usuário, o quanto cada um daqueles conceitos é similar ao conceito pesquisado, em cada um dos vocabulários. Desta forma, o sistema de recuperação da informação pode iniciar sua busca em uma das estruturas taxonômicas de um dos vocabulários participantes e, a partir daí, oferecer os mapeamentos para todos os outros vocabulários armazenados em seu espaço conceitual, fornecendo ao usuário as identidades descobertas e oferecendo as possibilidades semânticas a partir dos registros de conceito e seus mapeamentos.

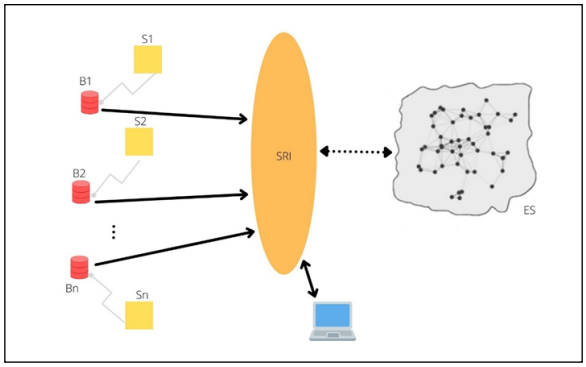

Figura 3. Sistema de recuperação da informação atuando no espaço semântico (Barbosa, 2021)

Um ambiente completo para representação deste sistema pode ser visto na figura 3, onde podemos ver os diversos participantes deste processo para a recuperação de informações distribuídas e indexadas por vocabulários heterogêneos.

Os diferentes passos para a criação e funcionamento deste sistema inteligente de recuperação são dados como:

1. O agente de software AS, programado com algoritmos que executam as técnicas demonstradas em nosso experimento (Barbosa, 2021) e em conformidade com a diretriz 5, executam continuamente seus códigos programáticos, extraindo as informações dos sistemas de organização do conhecimento S1, S2, ..., Sn, e criam e mantém atualizado o espaço semântico ES. Este processo é contínuo, pois uma vez gerado este espaço semântico, as atualizações e manutenções realizadas nos SOCs devem estar representados no espaço semântico. Esta base de dados comporta, portanto, os registros de conceitos extraídos dos SOC e seus interapontamentos, com suas distâncias semânticas calculadas. Em acordo com a diretriz 4, estes registros são identificados sob a forma de URI, garantindo a sua identificação única no sistema;
2. Um usuário interessado em recuperar informações das bases de dados indexadas pelos sistemas de organização do conhecimento do ambiente, faz acesso ao sistema de recuperação da informação SRI, que estabelece uma comunicação com o usuário, buscando no espaço semântico e mostrando os conceitos encontrados para cada SOC e suas distâncias semânticas, podendo oferecer ao usuário a escolha de um limite de similaridade que atenda sua busca. Consideramos que esta busca, conforme descrevemos e propusemos em nosso texto e reafirmamos aqui, não apresenta os mapeamentos matriz de duas dimensões, em linha-coluna, mas sim se apresentam em uma estrutura navegacional multidimensional, onde conceitos estão interligados e são apresentados como uma possibilidade de descobrimento de conhecimento por parte do usuário.
3. Uma vez que definidos os conceitos participantes da busca, em cada SOC participante, de forma automática ou com opcional interferência do usuário, o SRI, então extrai das bases de dados B1, B2, ..., Bn, os documentos indexados pelos termos representativos dos conceitos em cada base de dados. Reafirmando, portanto, que esta recuperação em cada base de dados pode ser feita com os termos sintáticos exatos buscados pelo usuário e confirmados pela similaridade semântica estabelecida pelos processos da diretriz 5, mas também pode ter expressão verbal diferente na recuperação dos documentos em cada base, uma vez que o conceito buscado pode ter esta representação diferente por diversos motivos em cada SOC participante, mas similaridade em seu significado semântico.
4. As escolhas dos limites semânticos estabelecidos e posteriormente validados pelo usuário em sua busca podem ficar armazenadas no sistema, de forma que componham um aprendizado para que futuras buscas multivocabulários possam utilizar este conhecimento acumulado.
5. A adição de novos sistemas de organização do conhecimento a um ambiente como este enriquece o espaço semântico gerado e permite um processo cada vez mais rico de recuperação de informações de forma inteligente, mesmo se tratando de bases de dados diferentes e heterogêneas.

Esta diretriz, portanto, se refere ao processo de construção de um sistema de recuperação da in-

formação que seja capaz de interpretar os objetos 'registros de conceito' armazenados no espaço semântico e fornecer ao usuário buscante uma informação recuperada com ótimos índices de precisão e revocação, superando a barreira da heterogeneidade e da simples igualdade sintática para sua solução.

Portando, como produto desta diretriz obtemos um ambiente de recuperação inteligente da informação que trabalha com os registros de conceitos, ou seja suas identidades semânticas, e suas distâncias semânticas. Este ambiente, composto centralmente pelo espaço semântico construído, é flexível em sua constituição, pois permite a inclusão de novos SOC, que indexem novas bases de dados e é aberto em sua concepção pois utiliza as tecnologias padronizadas e abertas da web semântica. Além, disso é um ambiente interativo, que propõe a participação do usuário ao permitir sua interferência no nível de compatibilidade pretendido.

4. Técnicas aplicadas

Consideramos que o caminho apontado pelas nossas diretrizes para compatibilização e correspondência de vocabulários heterogêneos, a partir de bases teóricas desenvolvidas por autores da Ciência da Informação, deve poder ser colocado em prática utilizando-se, entre outras, técnicas de manipulação de cadeias de caracteres e análises de hierarquias que sejam adequadas ao propósito desejado. Para isso recorremos a diferentes técnicas, algoritmos e sistemas já desenvolvidos pela Ciência da Computação para que seja possível implementar de forma prática nossa proposta teórica.

Inicialmente abordaremos algumas técnicas de correspondência que podem ser utilizadas por algoritmos e sistemas para identificar as possíveis correspondências entre termos de vocabulários. Conforme Euzenat e Shvaiko (2013) e Angermann e Ramzam (2017), e a partir da análise dos procedimentos apresentados em Achichi et al. (2017), Algergawy et al. (2018) e Algergawy et al. (2019), foi possível confirmar nove grandes grupos de tipos de técnicas de correspondência (cada um com diversas subcategorias e especificidades), em que cinco são voltadas para o nível de elemento, com valores literais, e quatro voltadas para o nível de estrutura usando uma estrutura "é-um".

As Técnicas de Nível de Elemento utilizam os valores literais dos conceitos, e/ou suas propriedades, para medir a similaridade semântica. Podemos preliminarmente citar cinco técnicas de nível de elemento: baseadas em recursos formais, baseadas em recursos informais, baseadas em strings (cadeias de caracteres), baseadas em linguagem e baseadas em restrições.

As técnicas baseadas em recursos formais se reportam e mapeiam a conhecimentos prévios fortemente estruturados. Estes recursos podem ser, por exemplo, taxonomias de mais alto nível, específicas para um domínio e padronizadas, como taxonomias que representam grupos de diferentes domínios ou taxonomias de mesmo domínio, mas mais gerais e abrangentes. As baseadas em recursos informais também usam a mesma técnica, mas podem se referir a recursos não padronizados, tais como diretórios de índices estruturados em nível superior às taxonomias a serem compatibilizadas. Nesses dois casos os elementos das taxonomias são apenas comparados e mapeados aos elementos das taxonomias globais.

Ainda no nível de elemento temos as técnicas baseadas em cadeias de caracteres que identificam correspondências com base na comparação e igualdade destas cadeias. Estas técnicas tratam de avaliar a comparação entre os termos e até de suas descrições (Cheatham e Hitzler, 2013). Esta similaridade pode ser calculada, de modo geral, de dois modos: a similaridade de nome e a similaridade de descrição.

A similaridade de nome mede quão similar uma palavra ou grupo de palavras é similar a outra ou a outras. Estas medidas podem ser feitas de múltiplas formas, conforme vemos a seguir. A distância Levenshtein define esta similaridade como o mínimo número de trocas necessário para transformar uma cadeia de caracteres em outra. Cada troca pode ser a transformação necessária para um caractere, seja a sua remoção, inserção ou substituição (Levenshtein, 1966). A distância de Bailey, denominada pelo autor como Euclidiana, mostra o comprimento da conexão necessária para combinar um ponto no espaço euclidiano com outro ponto. Por meio deste, cada caractere de uma sequência é atribuído a um ponto no espaço euclidiano (Bailey, 2004). Já o processo de distância de Hamming (Hamming, 1950; Tanenbaum, 2003) exige que as duas cadeias tenham o mesmo comprimento em número de caracteres e apresenta o número de caracteres diferentes em um mesmo índice posicional. A medida de distância de Lin (Kernighan e Lin, 1970), calcula a probabilidade de uma string ocorrer dentro de um termo. Por fim, Wu e Palmer (1994) classificam cada termo de acordo com a sua profundidade dentro de um corpo de texto usado para comparação.

A similaridade de descrição, por sua vez, leva em consideração termos compostos que devem ser comparados com outras sequências. As principais medidas de similaridade por descrição usadas

hoje são: a distância Jaccard, que representa a semelhança entre dois conjuntos de strings, a similaridade Cosine que considera as sequências como vetores para compará-las e a TF-IDF (Term Frequency–Inverse Document Frequency), que usa a importância de um termo, com base em sua ocorrência em um documento, como uma das bases de comparação (Jones, 1972; Tan et al. 2005).

As próximas técnicas do nível de elemento são as técnicas de correspondência baseadas em linguagem. Estas técnicas são normalmente usadas em conjunto com as técnicas baseadas em cadeias de caracteres e a comparação é geralmente apoiada pelo uso de conhecimento referente ao domínio, que permite analisar o contexto dos conceitos comparados. Podemos citar seis categorias predominantes nestas técnicas:

A Lematização e Morfologia agrupam diferentes formas de inflexão de uma palavra, de forma que elas possam ser analisadas como um único item, como por exemplo seu tipo mais comum, singulares e plurais. Tratam também de coisas tais como abreviaturas.

A Tokenização quebra um texto em palavras, frases, símbolos, ou outros "tokens" significantes. Um único token pode conter mais de uma palavra. Nesse caso o método é apelidado de N-Gram, onde N associa o número de palavras associado ao token.

Já a Eliminação reduz os tokens eliminando os elementos considerados supérfluos, por exemplo, stop-words.

Os métodos com léxicos ou tradutores são usados para traduzir entre idiomas. Normalmente são usados tradutores automáticos, tais como Microsoft Bing e Google Tradutor.

O método de similaridade é utilizado para analisar uma possível similaridade semântica entre conceitos, usando bases de dados, como, por exemplo, a WordNet.

Em seguida, temos a desambiguação de sentidos, que é utilizado para analisar o sentido da sentença no contexto considerado, ou seja, qual o token mais importante a ser considerado para comparação.

Por fim, para completar as técnicas de nível de elemento, temos as técnicas de correspondência baseadas em restrições, que analisam a estrutura interna do sistema de organização do conhecimento utilizado. Sempre agindo em conjunto com outros métodos, esta técnica pode avançar na superação da heterogeneidade conceitual. Podemos dividi-los em duas categorias. Consideramos inicialmente uma similaridade por tipo dos atributos, porque estes elementos descrevem os conceitos em um domínio. Nesse caso, dois conceitos de mesmo tipo, mas de nomes diferentes, que compartilhem a mesma descrição em diferentes atributos podem ser assumidos como similares semanticamente. Por exemplo, dois conceitos tais como "carro de passeio" e "automóvel" que estejam compartilhando atributos como número de portas, espaço na mala e número de bancos, podem ser assumidos como semanticamente similares. A outra categoria que temos é chamada de propriedades-chave, que são usadas para descrever os conceitos que pertencem, por exemplo, a uma taxonomia. Nesse caso, quando os conceitos dentro de uma taxonomia são estruturados de acordo com um determinado ponto de vista correspondente, as taxonomias como um todo podem ser assumidas como similares (Angermann e Ramzan, 2017).

Consideramos a seguir outro grupo de técnicas, aquelas baseadas em taxonomia e que se dedicam a explorar os subconceitos (especialização) e superconceitos (generalização) em uma taxonomia, também conhecidos com relações do tipo é-um. As taxonomias podem diferir no número total de conceitos e na quantidade de relações utilizadas. A análise de similaridade de dois conceitos, em diferentes estruturas, avalia seus subconceitos e seus conceitos superordenados e quanto menos diferentes estas estruturas forem, mais similar semanticamente eles serão. Estas técnicas, em adição às técnicas em nível de elemento, nos permitem criar algoritmos que estabeleçam as medidas semânticas propostas em nossas diretrizes.

Com as técnicas baseadas em grafos, uma taxonomia é considerada como um grafo identificado. Assim, as relações de paridade, ou irmandade, também são tomadas em consideração ao comparar conjuntos e subconjuntos e a distância entre cada um, usando técnicas matemáticas de análise de grafos, tais como, homomorfismo, similaridade de caminhos, similaridade de filhos e similaridade de folhas.

As próximas técnicas consideradas são bem interessantes e focam nas técnicas baseadas em instância. Neste caso a indicação de similaridade entre dois conceitos depende de suas instâncias. Esta similaridade, assim como as anteriores, também depende de dois conjuntos a serem comparados, pois define que conceitos similares devem ter instâncias similares. Apesar de nem todos os sistemas de organização do conhecimento apresentarem a ocorrência de instâncias em suas representações do conhecimento de um domínio, estes objetos, quando presentes, são

de grande utilidade na comparação e no estabelecimento de mapeamentos semânticos e medidas entre conceitos.

Por fim, de uso bastante limitado na literatura e de descrição bastante dispersa e pouco densa, temos as técnicas de correspondência baseadas em modelos que usam 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 e, Raciocínio de Lógicas de Descrição que é uma família de linguagens formais de representação do conhecimento. Um raciocínio é uma técnica que é capaz de inferir consequências lógicas de um conjunto de entidades (Angermann e Ramzan, 2017).

As técnicas e métodos mostrados acima são, na verdade, grupos e categorias de métodos, onde em cada categoria temos diversas variações e especificidades. Desde os mais complexos com técnicas de manipulação de grafos até aqueles de manipulação de cadeias de caracteres, são métodos não necessariamente criados para os propósitos de compatibilização, mas sim usados em diferentes aplicações e gerados por variados propósitos. A utilização destas técnicas pode e deve levar a serem combinadas de múltiplas formas ao serem aplicadas em um processo de compatibilização. Estas combinações de técnicas geram os diferentes algoritmos e procedimentos que podem ser utilizados sobre sistemas de organização do conhecimento para mapeamento e alinhamento de termos, visando a recuperação dos documentos e informações indexados por estes termos. Em suma, um algoritmo de correspondência usa uma estratégia peculiar consistindo em uma ou mais (geralmente mais de uma) técnicas de correspondência para superar a heterogeneidade de vocabulários. Estes algoritmos, conforme extraído dos relatórios recentes da OAEI (Ontology Alignment Evaluation Initiative), podem ser agrupados naqueles voltados para resolver quatro tipos de heterogeneidade, a saber, terminológica, conceitual, sintática e semiótica.

A heterogeneidade terminológica ocorre quando os descritores dos conceitos são diferentes, podendo ocorrer pelo uso de diferentes idiomas, por exemplo, ou diferentes sublinguagens técnicas, ou pelo uso de sinônimos. Em suma, pelo diferente uso do idioma.

A heterogeneidade conceitual ocorre quando duas taxonomias usam diferentes modelos, representando o domínio em questão com diferentes conceitos, por exemplo, dois conceitos semanticamente similares têm em uma taxonomia um número diferente de subconceitos em relação a outra taxonomia.

Já a heterogeneidade sintática com um viés estrutural, aqui neste caso, ocorre quando diferentes modelos de dados são utilizados para armazenar as taxonomias, por exemplo, uma armazenada em OWL e outra armazenada em RDF. Nesse caso, preliminarmente é necessário realizar uma tradução entre os formatos ou linguagens de representação.

A heterogeneidade semiótica surge quando pessoas fazem diferentes interpretações cognitivas dos conceitos, em especial, nas relações é-um. Por exemplo, quando um usuário de uma taxonomia não espera encontrar “Marcopolo” e “Ferrari” na mesma categoria de “Automóveis”, por exemplo, porque apesar de ambas serem marcas de tipos de veículos, uma serve como meio massivo de transporte de pessoas e o outro representa um modo de dirigir individual e esportivo.

Todas estas técnicas apresentadas aqui podem ser usadas para criar sistemas de correspondência, onde definimos que um sistema para correspondência de taxonomias e ontologias é um aplicativo ou um conjunto de aplicativos de software que tem por objetivo identificar e resolver diversos tipos de heterogeneidade na execução de uma operação de correspondência (Otero-Cerdeira et al., 2015). Ou seja, o que chamamos aqui de sistema de correspondência é um programa de computador desenvolvido para resolver um determinado problema de compatibilização entre ontologias específicas em um determinado contexto. Para isso o desenvolvedor utiliza bases de dados, ou data-sets, pré-determinados para aplicar diferentes algoritmos e gerar um mapeamento entre dois SOC em questão.

Portanto, a determinação da equivalência entre os conceitos e, mais importante, conforme descrito em nossas diretrizes, o estabelecimento das medidas de compatibilidade semântica entre dois conceitos pode ser alcançado com a aplicação dos métodos resumidamente descritos nesta seção, metodologicamente organizados em aplicativos de software escritos para este fim. Estes aplicativos de software, ao atuar sobre a nomenclatura dos conceitos, sobre sua posição relativa nas hierarquias taxonômicas, sobre a comparação de suas instâncias e sobre suas definições, quando presentes, são capazes de estabelecer as medidas de compatibilidade entre conceitos, conforme representado esquematicamente na figura 1. Desta forma a navegação neste espaço semântico criado será capaz de estabelecer uma compatibilização entre vocabulários heterogêneos de forma replicável, aberta e automática,

executada por máquinas, com vista a um processo de recuperação da informação inteligente e preciso.

5. Considerações finais

A linha de raciocínio que norteou o desenvolvimento desta pesquisa sempre foi o estudo e a compreensão das causas da heterogeneidade entre sistemas de organização do conhecimento e os caminhos para a sua interoperabilidade, de forma que seja possível estabelecer processos de recuperação inteligente da informação. Estes processos devem ter por objetivo permitir aos usuários recuperar informações de bases diversas indexadas por vocabulários heterogêneos, sem que seja necessário alterar estes vocabulários e sem que o usuário precise manualmente percorrer diferentes estruturas taxonômicas, dispendendo tempo precioso com informações imprecisas e com análise manual de múltiplas fontes e bases de dados. Em outras palavras, procuramos estudar e propor soluções que permitam que a capacidade de obter da Internet e da Web dados que respondam às necessidades dos indivíduos seja capaz de ser tão efetiva como o avanço do desempenho dos computadores e das redes.

Para isto este foi nosso foco principal: a recuperação da informação. Como diz Lévy, o cenário contemporâneo dificulta conseguir achar a informação que nos interessa e que tem mais valor, pois buscar e receber uma resposta avassaladora de informações não relevantes ou não significativas é quase tão ruim como buscar informações e conseguir achar pouco ou nenhum resultado.

Portanto, nossa pesquisa enfrenta na prática um problema que o mundo tem hoje, ao tentar produzir soluções para a recuperação da informação, não somente de dados de pesquisa, mas também bases bibliográficas, de livros e publicações científicas e dados em geral dispersos, armazenados e indexados de forma descentralizada. Isto nos levou a este trabalho, que se soma aos esforços de estudar e propor soluções de forma interdisciplinar que superem a heterogeneidade em todos os níveis dos dados digitais produzidos atualmente.

As propostas de Neville e Dahlberg têm destaque e importância reconhecidos na Ciência da Informação para a compreensão e solução do processo de compatibilização de vocabulários, mas tem uma grande dependência de atuação do ser humano. Assim, como seus próprios autores colocam, estes procedimentos não foram feitos diretamente para implementação em processos automatizados, mas como pudemos demonstrar em nosso experimento, tem um papel de grande importância com sua contribuição metodológica para dar forma e sentido à utilização de técnicas computacionais cujos propósitos se voltem para a compatibilização e correspondência de sistemas de organização do conhecimento.

Por fim, com base nos estudos e experimentos realizados e mostrados com detalhes em Barbosa (2021), fomos capazes de apontar um caminho com diretrizes que possam ser capazes de serem aplicadas a vocabulários de ambientes heterogêneos e gerar um espaço semântico. Possibilitando, assim, que a partir daí, possa ser utilizado por um SRI que permita a recuperação inteligente da informação nestes ambientes, de forma flexível, permitindo sua expansão, e interativa, que permita a participação do usuário de forma ativa nas suas buscas.

Portanto, nossa proposta foi apresentar aqui um caminho possível a ser implementado utilizando os recursos e a inteligência que já estão presentes nos sistemas de organização do conhecimento tradicionais que, aliados às técnicas e procedimentos computacionais que já temos disponíveis e às bases teóricas da Ciência da Informação, possa apontar uma solução para a heterogeneidade. Este caminho e diretrizes podem, por um lado, resolver problemas e compatibilidade com vocabulários já existentes, mas podem também serem usados para orientar e propor possibilidades de construção mais adequadas para SOC ainda a serem desenvolvidos, que utilizem novas tecnologias.

Referências

Achichi, M. et al.(2017). Results of the Ontology Alignment Evaluation Initiative 2017. Ontology Alignment Evaluation Initiative Conference. Viena.

Agraev, V. A. et al.(1974). Information retrieval system compatibility. // Automation Documentation and Mathematical Linguistics. 2, 29-37.

Algergawy, A. et al.(2018). Results of the Ontology Alignment Evaluation Initiative 2018. Ontology Alignment Evaluation Initiative Conference. Monterey.

Algergawy, A. et al.(2019). Results of the Ontology Alignment Evaluation Initiative 2019. Ontology Alignment Evaluation Initiative Conference. Auckland.

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. Proceedings of the 11th International Semantic Web Conference. Berlin, Germany: Springer. 394-408.

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

Bocatto, V. R. C.; Torquetti, M. C. (2012). Interoperabilidade entre linguagens de indexação. // Informação & Informação. Londrina. 17:3, 76-101.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. // Annual Review of Linguistics. 6, 213-234.

Campos, M. L. A.; Campos, M. L. M.; Davila, A. M. R.; Gomes, H. E.; Campos, L. M.; Lira, L. (2009). Aspectos Metodológicos no Reúso de Ontologias: um estudo a partir das anotações genômicas no domínio dos tripanosomatídeos. // RECIIS. Revista Eletrônica de Comunicação, Informação & Inovação em Saúde. 3, 64-75.

Cheatham, M.; Hitzler, P. (2013). String similarity metrics for ontology alignment. International Semantic Web Conference ISWC 2013. Heidelberg: Springer. 294–309.

Coates, E. J. (1970). Switching languages for indexing. // Journal of Documentation. London. 26:2, 102-110.

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

Euzenat, J.; Shvaiko, P. (2013). Ontology Matching, 2 ed. Heidelberg: Springer.

Gantz, J.; Reinsel D. (2010). The digital universe decade - are you ready? IDC White Paper. May.

Gardin, J. C. (1967). Recherches sur l'indexation automatique des documents scientifiques. // Revue d'informatique et de recherche opérationnelle. 1:6, 27-46.

Gardin, J. C. (1973). Document analysis and linguistic theory. // Journal of Documentation. London. 29:2, 137-68.

Gardin, N. (1969). Le lexique intermédiaire: un nouveau pas vers la cooperation internationale dans le domaine de l´information scientifique et technique. // Bulletin de l'UNESCO: à l' Intention des bibliothèques. Paris. 23:2, 66-71.

Goldberg, Y. (2017). Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies.

Hamming, R. W. (1950). Error detecting and error correcting codes. // Bell System Technical Journal. 29:2, 147-160.

Hammond, W; Rosenborg, S. (1962). Experimental study of convertibility between large technical indexing vocabularies. // Technical report IR-1. Datacontrol Corporation. Silver Spring. Ago.

Henderson, M. M. et al.(1966). Cooperation, Convertibility, Compatibility Among Information Systems: A Literature Review. // National Bureau of Standards. Jan.

Horsnell, V. (1975). The Intermediate Lexicon: an aid to international co-operation. // Aslib Proceedings. 27:2, 57-66.

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

Kernighan, B.; Lin, S. (1970). An efficient heuristic procedure for partitioning graphs. // Bell System Technical Journal. Blackwell Publishing. 49, 291.

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

Lévy, P. (2014). A Esfera Semântica. Tomo 1: Computação, cognição, economia da informação. Editora Annablume.

Lévy, P. (2009). From Social Computing to Reflexive Collective intelligence: The IEML Research Program. CRC, FRSC, University of Ottawa.

Lévy, P. (2019). IEML - A metalinguagem da Economia da Informação - Livro Branco. Pré-print, não publicado.

Newman, S. M. ed. (1965). Information Systems compatibility. Washington: Spartan Books.

Nurcan, S. et al.(1999). Change process modeling using the EKD – Change Management Method. 7th European Conference on Information Systems, ECIS' 99. Copenhagen, Denmark. 513-529.

Otero-Cerdeira, L.; Rodríguez-Martínez, F. J.; Gómez-Rodriguez, A. (2015). Ontology matching: a literature review. // Expert Systems with Applications. 42, 949.

Smith, L. C. (1974). Systematic searching of abstracts and indexes in interdisciplinary areas. // American Society of Information Science. 25, 343-353.

Soergel, D. (1972). A universal source thesaurus as a classification generator. // Journal of the American Society for Information Science. 23:5, 299-305.

Soergel, D. (1974). Indexing languages and thesauri: Construction and maintenance. Los Angeles, CA: Melville Wiley Information Science Series.

Statista (2021). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. // Statista Research Department. https://www.statista.com/statistics/871513/worldwide-data-created/ (2022-03-01).

Svenonius, E. (1975). Translation between hierarchical structures: an exercise in abstract classification. Ordering systems for global information networks. 204-211.

Tan, P. N.; Steinbach, M.; Kumar, V. (2005). Introduction to data mining. 1st. ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Tanenbaum, Andrew S. (2003). Redes de Computadores. 4ª Edição. Editora Elsevier.

Unesco (1971). Unisist study report on the feasibility of a world science information system. Paris.

Wellisch, H. (1972). A concordance between UDC and Thesaurus of engineering and scientific terms. Proceedings of the International Symposium UDC in Relation to Other Indexing Languages. Novi, Yugoslavia.

Wersig, G. (1975). Experiences in compatibility research in documentary languages: Ordering systems for global information networks. 423-430.

Wu, Z.; Palmer, M. (1994). Verbs semantics and lexical selection. 32nd Annual Meeting on Association for Computational Linguistics (ACM). New Mexico. 133-138.

Enviado: 2022-03-30. Segunda versión: 2022-09-26.
Aceptado: 2022-09-26.