XXII Encontro Nacional de Pesquisa em Ciência da Informação – XXII ENANCIB

ISSN 2177-3688

GT-8 – Informação e Tecnologia

RECUPERAÇÃO DA INFORMAÇÃO EM REPOSITÓRIOS DIGITAIS: UM ESTUDO SOBRE *ONTOLOGY LEARNING*

INFORMATION RETRIEAVAL IN DIGITAL REPOSITORIES: AN ONTOLOGY LEARNING STUDY

William Pires de Castro. UNESP.

José Eduardo Santarem Segundo. UNESP.

Modalidade: Resumo Expandido

Resumo: Os repositórios digitais atuam como bibliotecas *online*, gerindo grandes quantidades de dados, mas trazendo problemas no processo de recuperação da informação. Este trabalho objetiva entender definições e levantar a viabilidade de se aplicar processos de *Ontology Learning* em repositórios digitais, para garantir melhorias na recuperação da informação. Os procedimentos metodológicos foram a pesquisa exploratória e descritiva, a partir de um protocolo para levantamento dos dados. Conclui-se que o *Ontology Learning* pode trazer evolução para o cenário de ontologias e repositórios digitais, já que seu funcionamento foi comprovado em cenários diversificados e tem fácil implementação em bases de cunho textual.

Palavras-Chave: Repositórios Digitais. Ontologias. Ontologia Autogerenciada. Recuperação da Informação.

Abstract: *The digital repositories work like online libraries, managing big data contents, but bringing problems in the processes of information retrieval of itself. This work has the objective to understand definitions and check the viability to apply Ontology Learning process in digital repositories, to grant improvement in information retrieval process. The methodological procedures adopted was exploratory research and descriptive, from a protocol for data collection. It is concluded that the Ontology Learning process can bring evolution to the scenario of ontologies and digital repositories, since their operation was done together and easily implemented on a text basis*.

Keywords: Digital Repositories. Ontology. Ontology Learning. Information Retrieval.

1 INTRODUÇÃO

A Recuperação da Informação (RI) se trata do processo da extração de informação relevante de “documentos”, sejam eles digitais ou físicos.

Para este escopo, um “documento” pode ser definido pelas informações relevantes a respeito dele, como o nome, autor, qual a temática abordada pelo mesmo, entre outras coisas

que, deem relevância ao próprio, em determinado contexto, dando uso a informação presente nele.

Dentro de sistemas digitais de recuperação, buscas podem ser efetuadas por formas diversas, como palavras-chave, título e outras, que na maioria das vezes, não são capazes de representar a totalidade da informação de um documento. Dessa forma o conteúdo latente dos próprios documentos é deixado de lado para a realização de uma indexação mais rasa. A exploração do contexto dos conteúdos dos documentos pode trazer possibilidades promissoras, podendo ser uma solução futura para a quantidade massiva de documentos gerados pelas buscas.

Coneglian (2014) propõe que o uso de ontologias pode auxiliar no processo de recuperação da informação em repositórios digitais, melhorando a forma com a qual a informação é indexada. Entretanto, o processo de criação de uma ontologia por si só é extremamente trabalhoso, deixando o processo em alguns momentos muito custoso.

Ontologias podem ser criadas a partir de uma linguagem, que Zahra *et al.* (2014) citam ser a principal forma de se transferir conhecimento entre humanos, e são portadas para fontes textuais, que podem corroborar para o uso de ferramentas de aprendizagem de ontologias.

Uma vez que se facilita a criação e atualização de uma ontologia, também se viabiliza o processo de aplicação da própria em repositórios digitais.

Este artigo está caracterizado como uma pesquisa exploratória, descritiva qualitativa, a partir da análise de documentos referentes à temática.

A análise de documentos foi realizada de acordo com essas etapas: (1) Escolha das palavras-chave para nortear a pesquisa: Escolher palavras-chave que levem em consideração o resultado final, para garantir uma melhor recuperação da informação, diminuindo ao máximo a quantidade de documentos recuperados, com o objetivo de aumentar a performance da pesquisa. (2) Coleta de documentos para gerar conhecimento: coleta de documentos que pudessem agregar ao referencial teórico da pesquisa, como definições sobre ontologias, *ontology learning*, recuperação da informação, repositórios digitais, o uso de ontologias em repositórios digitais e o uso de *ontology learning* em ontologias aplicadas em repositórios digitais.

Para a validação do trabalho, o seguinte protocolo foi seguido:

Quadro 1 - Protocolo de busca da análise documental.

Protocolo de Pesquisa	
Pergunta da pesquisa	Existe algum trabalho de melhoria da recuperação da informação em repositórios digitais utilizando *ontology learning*?
Palavras-Chave	Ontologia, *Ontology* Learning, Repositórios Digitais, *Digital Repositories, Ontology*
Bases de dados consultadas	BRAPCI, Google Scholar, Repositório Unesp
Período Abrangido	indefinido
Idiomas	Português, Inglês
Período da coleta	Set/2021
Forma de análise dos documentos	Os documentos foram analisados com o objetivo de extrair informações relevantes e que pudessem se completar para construir uma narrativa sobre o *Ontology Learning*, repositórios digitais.

Fonte: Os Autores (2021).

Como resultados, a busca na BRAPCI se mostrou contundente quando focadas em ontologias, recuperação da informação e repositórios digitais, trazendo cerca de 1629 artigos respectivos a essas temáticas. Contudo, quando o enfoque é dado ao *ontology learning*, apenas 5 documentos são recuperados, criando assim a necessidade da expansão da base de dados para compor o referencial teórico da pesquisa.

Como detalhes a respeito do *Ontology Learning* (OL) precisavam ser respondidos, como trabalhos correlatos, ou usos que pudessem comprovar ou viabilizar a tentativa de uma aplicação em repositórios digitais.

A busca no repositório de teses da UNESP, veio para complementar com conteúdo nacional a respeito das temáticas, principalmente sobre o "uso de ontologias em repositórios digitais".

Diante disso, esta pesquisa tem como objetivo levantar a possibilidade do uso de OL em repositórios digitais baseados em ontologias, focando entender o processo de OL e sua aplicação em outros cenários.

2. RECUPERAÇÃO DA INFORMAÇÃO

Segundo Mooers (1951) o termo *"Information Retrieval"* (Recuperação da Informação) descreve tratativas de aspectos intelectuais da descrição da informação e sua especificação para busca, em conjunto com qualquer sistema, técnicas ou máquinas que são empregadas para realizar esta operação.

A recuperação da informação lida com a representação, armazenamento, organização e acesso às informações, provendo ao usuário aquilo que ele necessita de uma maneira facilitada (BAEZA-YATES E RIBEIRO-NETO, 2013).

O principal intuito da RI é atender as necessidades do usuário. Indicar o que seria mais ou menos relevante de acordo com o contexto da pesquisa direcionado a um conjunto de informações, ressaltando que, em alguns casos, nem o próprio usuário sabe o que está procurando (SANTAREM SEGUNDO, 2010).

Pode se afirmar que a área da RI foi criada para adquirir a informação mais adequada de maneira mais rápida, obtenção essa que se torna cada dia mais necessária, devido à alta quantidade de informação disponível por meio de Sistemas de Recuperação da Informação (SRI) abertos e privados, que são utilizados diariamente para diversos objetivos.

Vale salientar que a RI está relacionada, mesmo que indiretamente, com a representação, armazenamento, descrição, organização, preservação e acesso à informação. A representação e organização de itens de informação deveriam prover o uso, a preservação e o acesso à informação pelo interessado. Infelizmente, o acesso à informação necessária não é uma atividade simples (SANTAREM SEGUNDO, 2010).

Santarem Segundo (2017) comenta sobre essa crescente em outras áreas, como a *internet*, que detém o foco de muitas pesquisas envolvendo recuperação da informação, gerando conteúdo que precisa ser integrado por meio de relações semânticas a outros documentos e objetos digitais.

3 ONTOLOGIAS

Existem diversas definições para ontologia na literatura, definições dadas pela Ciência da Informação, Ciência da Computação e anteriormente dadas na antiguidade por Aristóteles.

O termo ontologia passou por diversas mudanças de significado, sendo o estudo da existência, representação de tudo que possa ser reconhecido para Leibniz, a objetividade das entidades lógicas entre a subjetividade e trocando a noção de termos lógicos como construções físicas já existentes.

Coneglian (2017) resume que o termo ontologia vem sendo usado com a característica de descrever o contexto de um domínio existente no mundo, sendo computacionalmente identificável no mundo real.

Em outras palavras, uma forma de representação de conhecimento dentro de um domínio/escopo de interesse, com suas características e seus relacionamentos.

Dentro do contexto da *Web* Semântica, as ontologias acabam auxiliando na comunicação entre as pessoas e a máquina, levando em consideração não apenas a sintaxe, mas a semântica do que está sendo prescrito.

As definições de Santarem Segundo (2010) e Campos e Campos (2014) permitem inferir que as ontologias podem ser utilizadas como vocabulários compartilhados, para diversos domínios, especialmente aqueles relacionados a iniciativas de dados abertos, que devem necessariamente possuir vocabulários compartilhados.

3.1 *ONTOLOGY LEARNING*

Para a produção de qualquer texto, é preciso que o autor aplique seus conhecimentos direta ou indiretamente a respeito do assunto. O processo de aprendizado de ontologia (*Ontology Learning*) consiste em que a partir de um texto apresentado, seja possível construir um domínio, em uma tentativa de abranger todo o conhecimento aplicado pelo autor.

Asim *et al.* (2018) alega que o desenvolvimento de ontologias trabalha com muitos campos do conhecimento, como processamento de linguagem natural, *machine learning*, *data mining.*

Ainda segundo o autor, o *data mining*, *machine learning* e a recuperação de informação providenciam técnicas estatísticas para extração de termos de domínio específico, conceitos e associações. O processamento de linguagem natural atua em cada nível da camada de *ontology learning* providenciando técnicas linguísticas.

O processo de aprendizagem de OL é formado por um conjunto de métodos e técnicas de construção semiautomática de novas ontologias ou para enriquecer ontologias já existentes (GÓMEZ-PÉREZ; MANZANO-MACHO, 2003). Todavia, para construir uma ontologia de maneira semiautomática, se faz necessário a automação do processo de aquisição de conhecimento e, para isso, diversas abordagens são sugeridas (MAEDCHE; STAAB, 2004).

Entre elas, Bedini e Nguye (2007) definem alguns elementos considerados cruciais, sendo eles: *Extraction, Analysis, Generation, Validion e Evolution.*

A fase de extração fica responsável por obter a informação necessária para se gerar a ontologia a partir de documentos existentes. Suas entradas podem ser realizadas por dados estruturados, semiestruturados e não estruturados, onde técnicas de processamento de linguagem natural, agrupamento, *machine learning*, semântica, morfológica ou léxica podem ser utilizadas, individualmente ou como se ocorre com mais frequência, combinadas.

A segunda fase, a análise, tem seu foco na combinação da informação recuperada, e em um possível alinhamento entre duas ou mais ontologias existentes, dependendo do caso. Esse passo precisa de técnicas já foram utilizadas no primeiro passo:

- Análise morfológica e análise léxica das camadas;
- Análise semântica para detectar sinônimos, homônimos e identificar atributos comuns;
- Técnicas baseadas em raciocinadores para detectar inconsistências e relações induzidas.

A geração (*Generation*) trabalha com a fusão de ontologias e da formalização do meta-modelo utilizado pela ferramenta em um formalismo mais geral, podendo ser interpretado por outras ferramentas, como OWL e RDF/S.

A validação (*Validation*) é a fase de validação automática dos resultados adquiridos, já que as outras fases podem gerar conceitos e relações erradas. Pode-se inserir uma validação ao final de cada passo anterior, tarefa que é realizada a mão, mas em alguns casos pode ser automatizada.

A fase de evolução (*Evolution*) avalia a capacidade das ferramentas de resolver o problema. É uma melhoria das aplicações em qualidade e número, onde pode existir alterações nas ontologias. Essa operação é considerada como uma soma de novos requisitos e pode demandar outra etapa de extração (*Extraction*) de informações para um novo alinhamento.

Zahra *et al.* (2014) complementa que nesse ponto, os avaliadores humanos ainda são responsáveis e necessários para incluir novos relacionamentos na ontologia.

3.2 USOS DA ONTOLOGY LEARNING

Atualmente existem vários estudos relacionados à criação de modelos de ontologia baseados em *ontology learning*.

No estudo "*A survey of ontology learning techniques and applications*" (2018), os autores condensam vários artigos com técnicas de OL em vários domínios.

Neste estudo os autores comentam que uma abordagem híbrida entre técnicas linguísticas e estatísticas podem produzir ontologias melhores. De acordo com os autores, depois de uma análise da literatura de *OL*, muitos pesquisadores preferem usar técnicas estatísticas, pois o desempenho das técnicas de OL acabam dependendo diretamente da eficiência do pré-processamento dos dados no domínio escolhido.

Asim *et al.* (2018) traz que os problemas ligados ao *ontology learning* vem da parte de processamento de linguagem natural, tendo muitos gargalos como a extração de textos não estruturados, correferências, reconhecimento de entidades e classificação de fala.

O autor também propõe diretivas para melhorar o processo de OL, como apresentado no quadro 2:

Quadro 2: Desafios *Ontology Learning*.

Desafio	Proposta
Diversidade na formatação dos dados, dados em línguas diferentes	Novas abordagens para integrar e harmonizar dados Algoritmos avançados de ontologias entre linguagens para aprendizagem de ontologias
Falta de validação automática de ontologias, ontologias defeituosas	Uso da web social, *tagging* colaborativa e folksonomia Uso de motores de busca para validação de resposta
Escalabilidade de técnicas de aprendizagem de ontologia	Aumento da pesquisa para acomodar conjuntos de dados maiores Organização de desafios da comunidade por órgãos governamentais para aumentar a escala de pesquisa de aprendizagem de ontologia
Requisito de intervenção humana para melhor qualidade das ontologias aprendidas	Necessidade de técnicas de pós-processamento automático Integrar framework de pós-processamento com framework de aprendizagem de ontologia para aumentar a qualidade da ontologia

	Uso de pesquisa nas áreas de *crowdsourcing* e jogos de computação baseados em humanos
Falta de ontologias robustas	Fortalecer algoritmos de aprendizagem de axioma

Fonte: traduzido de Asim (2018)

Outro trabalho relevante, é "*OntoSmart*, um modelo de recuperação da informação baseado em ontologias", onde Ferneda e Dias (2017) propõem um modelo onde a ontologia define os conceitos do vocabulário de domínio do *corpus* dos documentos, limitando o contexto onde a recuperação da informação será realizada, com uso de conceitos de indexação e expansão de consulta baseadas em ontologias, lidando com valor semântico e distância.

De acordo com os autores:

> antes de expressar sua necessidade de informação o usuário define o seu domínio de interesse por meio da seleção de uma ontologia, que será utilizada para agregar novos termos à expressão de busca inicialmente formulada por ele.

Dessa forma, este modelo apresenta uma alternativa para a recuperação da informação, a partir de uma base documental, conseguindo segregar em contexto, levando em consideração o conteúdo de cada texto.

Kadir, Aliane e Guessoum (2021) completam que muitos trabalhos remanescentes estudam como atingir sistemas de alta precisão, evitando intervenções humanas ao máximo, e que grande parte dos enfoques são muito dependentes da língua a qual é destinada, desfavorecendo a generalização de conhecimento específico.

Em resumo, o processo de OL precisa de uma base documental muito bem definida, para conseguir trabalhar com seu processo de adição ou remoção de conteúdo da ontologia, base esta que pode ser facilmente encontrada em repositórios digitais, já que ano a ano, vem tendo um aumento documental, possibilitando gerar ontologias mais eficientes.

3 CONSIDERAÇÕES FINAIS

Este artigo traz algumas definições sobre ontologia e *ontology learning* com a proposta de trazer uma discussão a respeito de seu uso em repositórios digitais, e como isso poder auxiliar na recuperação da informação nesses ambientes.

Acompanhando avanços contemporâneos, os repositórios digitais agregam com a pesquisa científica de forma satisfatória, sendo objeto de pesquisa e de recuperação da informação para pesquisadores da Ciência da Informação.

Pode-se concluir que, ao se ter uma forma de se criar e auto alimentar ontologias, utilizá-las no processo de recuperação da informação se torna algo muito mais convidativo para pesquisadores em questão, com soma de que a utilização de bases textuais são ideais para a implementação da OL, corroborando com sua aplicação em repositórios digitais.

Os passos requeridos para a construção do modelo de OL também são facilitados, já que, para que o modelo funcione perfeitamente em um ambiente de repositórios a ontologia deve sofrer alterações todas as vezes em que um novo documento é adicionado, alimentando assim a relação entre os documentos a partir de seu *corpus*, permitindo a existência de ontologias mais robustas, que por sua vez podem propiciar melhorias na recuperação da informação nesse escopo.

Como trabalhos futuros, segue a sugestão de uma implementação real de um sistema de RI focado em repositórios, utilizando uma ontologia, com possibilidade de se realizar buscas textuais, para haver um cenário propício para a aplicação do OL.

REFERÊNCIAS

ASIM, MUHAMMAD NABEEL; WASIM, MUHAMMAD; KHAN, MUHAMMAD USMAN GHANI; MAHMOOD, WAQAR; ABBASI, HAFIZA MAHNOOR. **A survey of ontology learning techniques.** Disponível em: https://academic.oup.com/database/article-abstract/doi/10.1093/database/bay101/5116160. Acesso em: maio 22

BEDINI, Ivan. Nguyen, Benjamin. **Automatic Ontology Generation: State of the Art.** Acesso em: maio 22

CONEGLIAN**,** Caio**. MODELO COMPUTACIONAL DE RECUPERAÇÃO DA INFORMAÇÃO PARA REPOSITÓRIOS DIGITAIS UTILIZANDO ONTOLOGIAS.** Disponível em https://repositorio.unesp.br/bitstream/handle/11449/148996/coneglian_cs_me_mar.pdf?sequence=3&isAllowed=y. Acesso em: maio/2021.

JESUS, Ananda. RECOMENDAÇÕES TEÓRICO-METODOLÓGICAS PARA A PUBLICAÇÃO DE DADOS BIBLIOGRÁFICOS ABERTOS E CONECTADOS. Disponível em: https://repositorio.ufscar.br/bitstream/handle/ufscar/14228/Dissertação_AnandaFernandadeJesus.pdf?sequence=5&isAllowed=y. Acesso em: set. 22

FERNEDA, Edberto. DIAS, Guilherme Ataíde**. OntoSmart: um modelo de recuperação de informação baseado em ontologia.** Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2081/1882. Acesso em: maio de 22

FURGERI , Sérgio. **ONTOART ONTOLOGIA EM XML PARA DESCRIÇÃO DE ARTIGOS.** Disponível em: https://brapci.inf.br/index.php/res/download/56158. Acesso em: set. 2022
MOOERS, C. Zatocoding Applied to mechanical organization of knowledge. **American Documentation**. Washington, v. 2, n. 1m p-20-32. 1951.

SANTAREM SEGUNDO, José Eduardo**. O uso de elementos semânticos no processo de recuperação da informação em ambientes digitais.** Disponível em: https://periodicos.ufsc.br/index.php/textodigital/article/download/1807-9288.2017v13n2p93/35678. Acesso em: maio de 22.

SANTAREM SEGUNDO, José Eduardo. **Representação iterativa: um modelo para repositórios digitais**. 2010. 224 f. Tese (Doutorado) - Universidade Estadual Paulista, Faculdade de Filosofia e Ciências, 2010. Disponível em: http://hdl.handle.net/11449/103346. Acesso em: abr. de 2022.