Estudos Ontológicos Aplicado ao Contexto: Base de Dados em Ciência da Informação - BRAPCI

Liliane Simões dos Santos[1] and Cláudio Gottschalg Duque[1]

[1] *Universidade de Brasília, Campus Universitário Darcy Ribeiro, Brasília, DF, Brasil*

Abstract

The Big Data phenomenon highlights the need to organize and represent information in digital environments. The semantic web is an approach that adds meaning and context to content indexed on the Internet and aims to improve the processes of organization, representation and retrieval of information in digital environments. The use of ontology favors the organization, representation and retrieval schemes of information in the semantic web, as it allows the creation of conceptual models that describe relationships and properties of objects and entities linked to a given domain, which facilitates machine learning processes and interoperability between information systems. The article intends to identify the contributions of applied ontology in the context of information science. Data Mining addressed the ontological studies indexed in the information science database - Brapci. When considering the descriptor "ontology" and the time interval "1972-2023", the search resulted in 332 articles. The data will be indexed in "xls" format. Metadata will be used to define categories. "Wordsmith Tools" will be used to explore the data considering semantic agreement, keywords, frequencies of use and occurrences, patterns of use and context. The data will be transformed into relevant information for the academic community through the Data Mining technique and will present indicators that will be used to elaborate the theoretical framework and develop the applied ontological model. The article demonstrates thematic relevance and presents elements that contribute to the area of knowledge in the scope of organization, representation and retrieval of information in digital environments.

Resumo

O fenômeno *Big Data* evidencia a necessidade de organização, representação e recuperação da informação em ambientes digitais. A *web* semântica é uma abordagem que adiciona significado e contexto ao conteúdo indexado na *internet* e visa melhorar os processos de gestão da informação no contexto digital. O uso de ontologia favorece os esquemas de organização, representação e recuperação da informação na *web* semântica pois permite criar modelos conceituais que descrevem relações e propriedades dos objetos e entidades vinculadas ao domínio determinado que facilita o processo de aprendizado de máquina e a interoperabilidade dos sistemas de informação. O artigo pretende identificar as contribuições da ontologia aplicada ao contexto da ciência da informação. A técnica de *Data Mining* aborda os estudos ontológicos indexados na base de dados em ciência da informação - BRAPCI. Ao considerar o descritor "ontologia" e o intervalo temporal "1972-2023" a busca resulta em 332 artigos. Os dados estão indexados em formato "xls". Os metadados são utilizados para definir categorias. O "*Wordsmith Tools*" é adotado para explorar os dados considerando a concordância semântica, as palavras-chave, as frequências de uso e ocorrências, os padrões de uso e contexto. Os dados são transformados em informação relevante para a comunidade acadêmica através de *Data Mining* e apresentam indicadores que são utilizados para elaborar o referencial teórico e desenvolver o modelo ontológico aplicado. O artigo demonstra relevância temática e apresenta elementos que contribuem com a área do conhecimento na perspectiva de gestão da informação digital.

Keywords

BRAPCI, Data Mining, Ontology, Web Semantic, Wordsmith Tools

Proceedings of the XVI Seminar on Ontology Research in Brazil (ONTOBRAS 2023) and VII Doctoral and Masters Consortium on Ontologies (WTDO 2023), Brasilia, Brazil, August 28 - September 01, 2023.

EMAIL: eipsimoes@gmail.com (L. S. Santos); klauss@unb.br (C. G. Duque)
ORCID: 0009-0004-3352-7375 (L. S. Santos); 0000-0003-3558-466X (C. G. Duque)

1. Introdução

O fenômeno *Big Data* evidencia a necessidade de gestão dos espaços informacionais em ambientes digitais. O volume expressivo de dados disponíveis em rede, a velocidade e a variedade de dados demonstram a necessidade de organização, representação e recuperação da informação a fim de mediar conflitos existentes no processo de indexação e recuperação da informação digital.

O conceito sociedade em rede é definido como espaço democrático do conhecimento. Neste contexto os dados atuam como artefato de integração para organizar, representar e recuperar a informação digital. As tecnologias e os sistemas de informação atuam como elemento estratégico que agrega valor ao otimizar os produtos e serviços de informação e abordar o uso de dados estruturados em sistemas integrados e dinâmicos que adotam o fluxo da informação para solucionar conflitos de inconsistência.

A *web* semântica é uma abordagem que adiciona significado e contexto ao conteúdo indexado na internet e visa melhorar os processos de gestão da informação digital. É definida como extensão da *World Wide Web* (*internet*). O artigo apresenta a técnica de *Data Mining* aplicada na base de dados em ciência da informação (BRAPCI).

O objetivo do artigo é identificar as contribuições da ontologia aplicada ao contexto da ciência da informação e aborda os estudos ontológicos indexados na BRAPCI. Considerou-se o descritor "ontologia" e o intervalo temporal "1972-2023". O *corpus* representa 332 artigos indexados na *web* semântica.

Os dados extraídos estão indexados em formato "xls". Os metadados são utilizados para definir categorias e orientar a criação de indicadores. Os dados são convertidos em informação relevante para a comunidade acadêmica e apresentam indicadores que são utilizados para elaborar o referencial teórico e desenvolver o modelo ontológico aplicado.

O "*Wordsmith Tools*" é um conjunto de *software* utilizado para análise e processamento de dados textuais. A ferramenta inclui análise de concordância, extração de palavras-chave, análise de frequência de palavras, análise de colocações, construção de listas de vocabulário, análise de n-gramas e análise de frequência de letras. A ferramenta é utilizada para explorar o *corpus* e considera a concordância semântica, as palavras-chave, as frequências de uso, as ocorrências, os padrões de uso e o contexto da amostra.

As ontologias definem conceitos e relações no escopo de domínio e os modelos conceituais estabelecem hierarquias de classes, propriedades e instâncias para descrever e indexar objetos digitais. A representação do conhecimento, a busca e o processo de recuperação da informação, a interoperabilidade sistêmica, a integração de dados, o fluxo da informação, o raciocínio e a inferência do aprendizado de máquina são exemplos de aplicações baseadas em ontologias. Os estudos ontológicos aplicados ao contexto da ciência da informação contribuem com a melhoria dos sistemas de informação e atende demandas informacionais ao minimizar falhas e inconsistências do processo de organização, representação e recuperação da informação digital.

O artigo demonstra a relevância temática e apresenta elementos que contribuem com os aspectos de gestão da informação digital ao favorecer os esquemas de organização, representação e recuperação da informação pois permite criar modelos conceituais que descrevem relações e propriedades de objetos e entidades vinculadas ao domínio determinado que facilita o processo de aprendizado de máquina e a interoperabilidade dos sistemas de informação.

2. Embasamento teórico e conceitual

A ciência da informação, a computação e a linguística aplicada utilizam a linguagem semântica como modelo de organização, representação e recuperação da informação digital. A temática apresenta relevância no estudo de elementos que norteiam os processos de gestão da informação digital.

O processo de indexação ocorre com a linguagem semântica que agrega informações estruturadas aos dados através de métodos e técnicas que consideram as características do *corpus* referência para realizar inferências e automatizar o fluxo de informação e o aprendizado de máquina. O processamento, a categorização de dados e a definição de termos e conceitos contribuem com a melhoria do processo de organização, representação e recuperação da informação digital pois rotula os dados com linguagem estruturada que identifica padrões de uso e métricas relacionais para inferir o contexto e orientar os resultados de busca do usuário.

Ao adotar a gestão de dados a informação é convertida em artefato de valor que otimiza os sistemas de informação pois o processo de recuperação da informação considera o índice de precisão e revocação para orientar os resultados de busca e atender as demandas de informação do usuário. Para realizar o processamento de linguagem natural da amostra usa-se as ferramentas tecnológicas disponíveis no núcleo interinstitucional de línguística computacional (NILC).

A linguística computacional é uma abordagem que atua na gestão de dados pois identifica padrões de uso e analisa as conexões existentes no intuito de transformar os elementos de informação em dados estruturados que auxiliam a organização, representação e recuperação da informação digital.

A técnica de *Data Mining* auxilia o processo de gestão da informação pois traz representatividade e intencionalidade ao processo de gestão da informação digital. Os signos linguísticos são elementos que organizam e representam a informação. Ao considerar a perspectiva dos sistemas de informação eles atuam na construção do conceito e definição da linguagem semântica durante o processo de indexação e construção do vocabulário controlado.

Borko 1968

Define ciência da informação como disciplina que investiga os modos de processamento da informação para otimizar o acesso e a usabilidade em sistemas de informação.

Saussure 1977

Postula a semiótica e aborda questões relacionadas a tríade: signo, significado e significante durante o processo conceitual e a definição de objetos no processo de comunicação.

Dahlberg 1978

Apresenta a teoria analítica do conceito onde o termo é associado a ideia de conceito e elemento linguístico. O conceito é definido como unidade do conhecimento representado por elementos de referência e evidências amostrais de conteúdo especializado. A representação do conhecimento é descrita como processo cognitivo que possui características padronizadas que considera a precisão de normas e definições conceituais.

Os elementos são atributos da linguagem de especialidade e os termos são artefatos para indexar conteúdo descritivo na *web* semântica. A descrição de conteúdo especializado adota a linguagem semântica e orienta a construção do conceito através da terminologia para organizar, representar e recuperar informações digital em relação as tipologias, funções, relações e extensões que moldam o conceito do objeto.

Wüster 1979

Propõe a teoria geral da terminologia (TGT) e estabelece a escola de Viena que é considerada pioneira nos trabalhos terminológicos. A TGT adota uma perspectiva prescritiva normativa que rotula e padroniza os termos e a definição de conceitos. Ela não considera as variações do contexto informacional.

O conceito é definido como unidade de pensamento e possui fundamentação normativa na construção da estrutura conceitual dentro da perspectiva cognitiva e compõe uma estrutura rígida e fixa que não aceita termos polissêmicos, sinônimos e homônimos que resulta em sistemas de informação com índice de precisão nas inferências orientadas ao aprendizado de máquina.

A representação do conceito é desenhada com base na terminologia que organiza a informação e representa o conhecimento estruturado através da linguagem de especialidade e a inferência do aprendizado de máquina. A terminologia atua como artefato de gestão da informação digital ao estudar os termos de uma área de especialidade.

Chomsky 1986

Aborda a ciência transformacional que adota a construção do conceito através da semântica como meio de representação do sentido ao utilizar a linguagem terminológica e o vocabulário controlado como artefatos de indexação.

Belkin 1990

Apresenta a abordagem cognitiva de ciência da informação. A representação do conhecimento é interdisciplinar pois dialoga com diversas áreas do conhecimento como a filosofia, a ciência da informação, a computação e a linguística.

Buckland 1991

Trata o conceito de informação orientado a coisas e objetos. Define a terminologia como ciência autônoma que estuda o processo de comunicação como elemento de expressão do conhecimento especializado através de métodos de análise comportamental e contextual dos termos. Denomina os sistemas de informação como unidade que coleta, trata, organiza e disponibiliza objetos informativos em ambientes digitais.

Wersig 1993

Menciona os sistemas de informação como atividades que realizam operações variadas e complexas de processamento e indexação de objetos digitais. Considera o armazenamento, o tratamento e a recuperação de dados sistematicamente com a finalidade de prover acesso ao usuário e promover a difusão do conhecimento.

Cabré 1995

Apresenta a teoria comunicativa da terminologia (TCT) que estuda conceitos e propõe a teoria interdisciplinar que engloba aspectos da teoria do conhecimento, da comunicação e de linguagem. As tecnologias atuam como recurso e solução. Define terminologia enquanto disciplina como conjunto de diretrizes ou princípios referente ao processamento e compilação de termos ou na perspectiva de produto ou serviço de informação ao representar uma coleção especializada de dados.

Sperber e Wilson 1995

Apresentam a teoria da relevância que aborda as expectativas de relevância aplicada ao contexto cognitivo de suposições. Apontam a precisão, a previsibilidade e as equivalências como *inputs* que atuam na organização, representação e recuperação da informação ao codificar e decodificar os enunciados. A recuperação da informação ocorre ao relacionar o contexto de suposições das inferências do processo dedutivo. A teoria da relevância adota critérios de especificidades para rotular dados e estruturar a informação digital.

Mcgee e Prusak 1998

Abordam o uso da informação orientada para estrutura organizacional. A informação exerce valor estratégico orientado para gestão por resultados e atua como ferramenta gerencial no contexto institucional. Mencionam conceitos de arquitetura da informação e a necessidade de compreender os papéis dos atores organizacionais durante o processo de organização, representação e recuperação da informação.

Davenport 2001

Adota o uso das ferramentas tecnológicas como artefato para implementar a gestão da informação digital ao organizar, representar e recuperar os dados.

Lancaster 2004

Menciona o impacto tecnológico na otimização dos serviços e produtos terminológicos. Cita o índice de revocação e precisão para mensurar a qualidade das inferências e o aprendizado de máquina nos processos de organização, representação e recuperação da informação digital. Considera a qualidade da indexação semântica para mensurar o processo de gestão da informação digital.

Le Coadic 2004

Define a ciência da informação como meio de analisar os processos de uso e gestão da informação digital. Define os sistemas de informação como produtos de informação que permitem a comunicação, o armazenamento e o uso compartilhado de dados.

Castells 2005

Aborda a sociedade em rede como espaço democrático do conhecimento. A globalização e as tecnologias servem como pontos de conexão e integração sistêmica (teia global). As tecnologias assumem papel ativo na gestão da informação digital ao organizar, representar e recuperar a informação digital. Menciona os espaços informacionais e o fluxo de informação como aplicação otimizada dos produtos e serviços da web semântica.

Robredo 2005

Define ontologia como esquema conceitual sobre determinado domínio que visa promover e facilitar a interoperabilidade dos sistemas de informação e otimizar o fluxo de informação ao integrar sistemas complexos que adotam as tecnologias para inferir e dimensionar resultados por analogias inteligentes. Os modelos relacionais e os princípios da hierarquia de conceitos atuam como modelos de organização, representação e recuperação da informação digital. A informação é definida como registro codificado que envolve variados formatos e suportes categorizados e indexados de modo estruturado.

Capurro 2007

Menciona o processo de assimilação do conhecimento como artefato de organização, representação e recuperação da informação através da indexação dos objetos digitais.

Gruber 2009

Define ontologia como artefato de gestão da informação que sistematiza e representa o conteúdo descritivo. É o modelo de representação utilizado na *web* semântica. Os padrões semânticos representam a linguagem que especifica conceitos e automatiza o fluxo de informação. A ontologia é definida como conjunto de termos do vocabulário controlado na relação de organização, representação e recuperação da informação digital.

3. Discussão e perspectivas futuras

No contexto digital é importante estruturar dados e categorizar a informação visando satisfazer a demanda informacional do usuário de informação ao organizar, representar e recuperar a informação digital. As inconsistências dos sistemas de informação representam fragilidades do processo de indexação. Geralmente, os dados não são indexados corretamente ou não adotam gestão qualitativa dos dados.

A experiência do usuário ao buscar informação em ambientes digitais deve orientar o processo de gestão da informação digital. A ausência de gestão qualitativa dos dados resulta em inconsistências geralmente ocasionadas por imprecisão terminológica ou ambiguidades conceituais no processo de inferência e aprendizado de máquina ao recuperar a informação digital. Ao estruturar os dados com linguagem semântica o objeto digital é indexado com elementos que auxiliam o processo de organização, representação e recuperação da informação digital além de otimizar os produtos e serviços de informação.

A linguagem semântica atua como modelo de gestão da informação ao organizar, representar e recuperar a informação digital e abordar o conhecimento através da indexação de objetos digitais. A relação de inferência, o processamento de dados, a categorização de dados e a assimilação de significados ocorre através das relações lógicas e semânticas entre termos e conceitos.

A temática promove melhorias e impactos qualitativos no processo de gestão da informação digital. O artigo abrange uma demanda interdisciplinar entre as áreas do conhecimento. Em especial, da filosofia, da ciência da informação, da linguística aplicada e da computação no cumprimento do papel normativo e social da informação nos aspectos de organização, representação e recuperação da informação digital.

A descrição com enfoque estatístico e qualitativo da amostra representada pelo *corpus* da BRAPCI é realizada por estratos analíticos que representam o índice de relevância, os parâmetros de frequência e as especificidades dos termos indexados. O índice de mensuração e satisfação adotou critérios de revocação e precisão.

O corpus é uma amostragem representativa do conteúdo descritivo composto por artigos indexados na base de dados em ciência da informação (BRAPCI). A criação do *corpus* adota

critérios de padronização terminológica. Adota-se o software "*Wordsmith Tools*" para explorar os dados considerando a concordância semântica, as palavras-chave, as frequências de uso e ocorrências, os padrões de uso e contexto. Os dados são convertidos em informação relevante para a comunidade acadêmica através da técnica de *Data Mining* e apresentam indicadores que são utilizados para elaborar o referencial teórico e desenvolver o modelo ontológico aplicado.

O artigo demonstra relevância temática e apresenta elementos que contribuem com a área do conhecimento em gestão da informação digital ao destacar funções, eliminar ambiguidades, controlar sinônimos e listar equivalências ao organizar, representar e recuperar as informações digitais e otimizar o fluxo de informação. A pesquisa em andamento vai incluir autores atuais que abordam a perspectiva do processo de recuperação da informação. O diálogo interdisciplinar pretende integrar a compreensão dos aspectos de organização, representação e recuperação da informação digital.

4. Agradecimentos

O estudo é realizado graças aos esforços de uma política pública de promoção da ciência e educação brasileira no cumprimento do papel social e normativo em prol da valorização e difusão do espaço democrático do conhecimento. Na intenção de reconhecer os papéis fundamentais de cada ator em suas respectivas atribuições que contribuíram significativamente com o desenvolvimento da pesquisa apresentada, agradeço:

À Universidade de Brasília (UnB) por se comprometer com a missão de inovar e incluir através das ações de ensino, pesquisa e extensão a fim de formar cidadãos qualificados para o exercício profissional e empenhados na busca de soluções democráticas e atuação de excelência para atender e integrar as demandas da sociedade.

Ao Programa de Pós-Graduação em Ciência da Informação (PPGCINF) e a Faculdade de Ciência da Informação (FCI) pela oportunidade de desenvolver competências científicas, aprofundar os conhecimentos adquiridos e aperfeiçoar a capacidade profissional no desenvolvimento de pesquisas em ciência da informação.

Ao Prof. Dr. Cláudio Gottschalg Duque (co-autor) pela orientação, disponibilidade e motivação durante a jornada acadêmica.

Ao Seminário de Pesquisa em Ontologias no Brasil (ONTOBRAS) e ao *Workshop* de Teses e Dissertações em Ontologias (WTDO) pela oportunidade e o espaço de discussão da proposta apresentada e o aprendizado através da rede de contatos e o compartilhamento de conhecimento com pesquisadores da área que contribuiu positivamente com a pesquisa em andamento.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio e recurso financeiro através do código de financiamento 001.

Referências

[1] Borko, H. Information science: What is it? American documentation, v. 19, n.1,1968.
[2] Saussure, F. 1977. Curso de lingüística geral. São Paulo: Cultrix/USP, 1977.
[3] Dahlberg, I. Teoria do conceito. Ciência da Informação, Rio de Janeiro, v.7, n.2, p.101-107,1978.
[4] Wüster, E. Introdução à teoria geral da terminologia e lexicografia terminológica. Springer, Viena 1979.
[5] Chomsky, N. Knowledge of language: its origin, nature and use. New York: Praeger, 1986.
[6] Belkin, N. J. The cognitive viewpoint in information science. Journal of information science, n.16. p.11-15. 1990.
[7] Buckland, M. Information as thing. Journal of the american society for information science. v.42, n.5, p.351-360, 1991.
[8] Wersig, G. Information science: the study of postmodern knowledge usage. Information processing & management, Oxford, U.K. v. 29, p. 229-239, Mar. 1993.

[9] Cabré, M. T. La terminologia hoy: concepciones, tendencias y aplicaciones. Ciência da informação, Brasília, v.24, n.3, p.289-298, set./dez. 1995.
[10] Sperber, D; Wilson, D. Relevance: communication and cognition. Cambridge/MA. Blackwell, 1995.
[11] Mcgee, J.; Prusak, L. Gerenciamento estratégico da informação. Rio de Janeiro: Editora Campus, 1998.
[12] Davenport, T. H. Ecologia da informação. São Paulo: Futura, 2001.
[13] Lancaster, F. W. Indexação e resumos: teoria e prática. 2. ed. Brasília: Briquet de Lemos, 2004.
[14] Le Coadic, Y. F. A ciência da informação. 2. ed. Brasília: Briquet de Lemos, 2004.
[15] Castells, M. A sociedade em rede. 8 ed. São Paulo: Paz e Terra, 2005.
[16] Robredo, J. Documentação de hoje e de amanhã: uma abordagem revisitada e contemporânea da ciência da informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília: Reproart, 2005.
[17] Capurro, R.; Hjørland, B. O conceito de informação. Perspectivas em ciência da informação, Belo Horizonte, v. 12, n. 1, p. 148-207, abr. 2007.
[18] Gruber, T. R. Ontology. 2009.
[19] Bufrem, L. S; Costa, F. D. O; Gabriel Junior, R. F; Pinto, J. S. P. BRAPCI: modelizando práticas para a socialização de informações: a construção de saberes no ensino superior. Perspectivas em ciência da informação, v. 15, n. 2, 2010.
[20] NILC - Núcleo interinstitucional de linguística computacional. Ferramentas de TIC'S. 1993.
[21] Worldsmith Tools. Scott, M. Wordsmith Tools, v. 6. Oxford, Oxford University Press. 2015.