Modelo de Arquitetura da Informação para *Ontology Learning* no ambiente de Nota Fiscal de Consumidor Eletrônica (NFC-*e*)

Resumo

A *Ontology Learning* trabalha em conjunto algoritmos de Aprendizado de Máquina, Mineração de Dado e Mineração de Texto e fornece um método poderoso para representar, reutilizar e compartilhar conhecimento de domínio. Desenvolver ontologias é uma tarefa difícil do ponto de vista da precisão, do tempo e do esforço, além de envolver uma grande subjetividade dos especialistas. Neste artigo, tem-se como objetivo apresentar e discutir a avaliação dos resultados da Arquitetura da Informação utilizada para organizar dados que serão recuperados por Mineração de Texto e utilizados por Aprendizado de Máquina na construção de uma ontologia sobre o produto cerveja. A revisão da literatura abordará temas como: Aprendizado de Máquina, Arquitetura da Informação, Ontologia e *Ontology Learning*, um resumo do método e da avaliação da organização da informação para engenharia de ontologia usando Mineração de Termos no contexto de Notas Fiscais de Consumidor Eletrônicas (NFC-*e*). A pesquisa inicia a partir de um planejamento para o levantamento bibliográfico e análise de exemplos que tratem de modelos de resolução adotados. O levantamento bibliográfico exploratório elabora uma revisão de literatura dentro dos termos no contexto da pesquisa, cujas fontes de informação utilizadas foram livros, artigos, periódicos, teses e dissertações. A metodologia adotada para obter os dados se baseou na primeira camada de 'Mineração do Termo' para *Ontology Learning* que apresentou uma resposta de cada elemento da AI como uma orientação para construção da ontologia, sem a necessária presença do especialista do domínio.

Palavras-chave: Arquitetura da Informação; Mineração de Texto; Aprendizado de Máquina; *Ontology Learning*; Notas Fiscais de Consumidor Eletrônicas (NFC-*e*).

1. Introdução

A inteligência artificial (IA) já é parte do cotidiano para solução de problemas reais. É uma evolução dos programas computacionais que, anos atrás, ainda tratavam por códigos as regras lógicas extraídas do conhecimento dos especialistas. O conhecimento, por sua vez, era extraído em entrevistas e se enfrentavam dificuldades pela subjetividade humana e pela pouca cooperação por parte do especialista.

A complexidade computacional, volume de dados, formatos diferentes de dados e o conhecimento espalhado em textos livres em diversos meios de

armazenamento motivaram o desenvolvimento de ferramentas computacionais sofisticadas e independentes da intervenção humana na área da IA, com a contribuição de sua subárea como Aprendizado de Máquina (AM), transformando a forma de interação com o computador com linguagens naturais, reconhecimento facial, visão computacional e comportamento inteligente a partir das escolhas do usuário. Isso propicia agilidade na entrega de sistemas computacionais avançados.

A ontologia promove um meio técnico adequado para compartilhar e trocar conhecimento entre humanos e/ou máquinas, “melhorando as capacidades de raciocínio e a compreensão da máquina” (DU, *et al*., 2024). *Ontology Learning* é uma área fundamental dentro deste domínio, encarregada da extração, da representação e do refinamento automático de conhecimento conceitual para a engenharia de ontologias. Utiliza-se para isso metodologias de IA como Mineração de Texto, Aprendizado de Máquina e Processamento da Linguagem Natural.

Utilizar o modo de AM para explorar a relação entre um objeto e suas propriedades e comportamentos enfrenta dificuldades, principalmente na configuração da IA e na seleção e organização da informação que será processada para o aprendizado e para a adaptação do ambiente. No entanto, a Arquitetura da Informação (AI) oferece requisitos que podem direcionar a escolha, a seleção e o tratamento de dados, facilitando a obtenção de conhecimento conceitual para a engenharia de ontologia por meio da orientação na descoberta automática de dados em textos livres e que ainda não foram totalmente exploradas nos estudos envolvendo AM.

Várias metodologias de desenvolvimento de ambientes de aprendizado de ontologias já foram propostas, porém não há espaço para um estudo prévio, anterior ao desenvolvimento da ontologia, que seja aprimorado para organizar dados de forma que se tornem requisitos da boa recuperação, com eficiência, segurança e rapidez, evitando a duplicidade de informações ou mesmo um grande esforço de mineração em informações desnecessárias.

Nesse contexto, Requisitos de Arquitetura da Informação inseridos de forma iterativa e incremental às camadas da *Ontology Learning* formam um modelo orientado para auxiliar a obtenção de termos, classes e relações, um artefato que melhora a modelagem e acelera a busca. Um cenário de Notas

Fiscais de Consumidor Eletrônica (NFC-*e*) é avaliado com base nos requisitos obtidos ao final da aplicação do modelo.

O objetivo deste artigo é apresentar e discutir a avaliação dos resultados da Arquitetura da Informação utilizada para organizar dados que serão recuperados por Mineração de Texto e utilizados por Aprendizado de Máquina na construção de uma ontologia sobre o produto cerveja.

A revisão da literatura abordará temas relevantes para a discussão proposta, quais sejam: Aprendizado de Máquina, Arquitetura da Informação, Ontologia e *Ontology Learning*, um resumo do método e da avaliação da organização da informação para engenharia de ontologia usando Mineração de Termos no contexto de Notas Fiscais de Consumidor Eletrônicas (NFC-*e*).

O trabalho propõe uma avaliação de informação coletada a partir de mineração de texto, usando elementos estruturais da Arquitetura da Informação e uma ontologia para auditoria em notas fiscais de consumidor eletrônica. A exploração dessa hipótese, de propor uma possível solução ao fato, com uso de arcabouços teóricos associados à Ciência da Informação, cria uma conexão de aproximação ou familiaridade com ele.

A pesquisa inicia a partir de um planejamento para o levantamento bibliográfico e análise de exemplos que tratem de modelos de resolução adotados. O levantamento bibliográfico exploratório elabora uma revisão de literatura dentro dos termos no contexto da pesquisa. As fontes de informação utilizadas foram livros, artigos, periódicos, teses e dissertações.

A metodologia adotada para obter os dados se baseou na primeira camada de ‘Mineração do Termo’ para *Ontology Learning* que apresentou uma resposta de cada elemento da AI como uma orientação para construção da ontologia, sem a necessária presença do especialista do domínio.

Para melhor organizar as informações, este trabalho se divide em revisão da literatura, trabalhos relacionados, desenvolvimento do modelo de Arquitetura da Informação para *Ontology Learning*, resultados e discussão e, por fim, as considerações finais.

2. Revisão da literatura

Arquitetura**,** como atividade milenar, se atém ao ordenamento das possibilidades humanas de construção, unindo, segundo Vitruvius, o belo

(*Venustas*), o útil (*Utilitas*) e a estrutura (*Firmitas*)" (KUROKI Jr., 2018), compatível a uma organização habilitada para recuperar aquilo que impacta a pessoa. Completa o autor que o termo arquitetar se associou ao termo informação, Arquitetura da Informação (AI), no sentido de submeter a informação a um ordenamento para melhor apreciação humana, de organizar decidindo, primeiramente, como se deseja procurar, achar algo na "construção da estrutura da informação que permite aos outros, posteriormente, entendê-la" (Wurman, 1997) de forma a conseguir recuperar a informação necessária e procurada. Para o autor, "[...] a estrutura da informação deve relacionar algo que já é compreensível para quem está sendo instruído de forma que seja possível a relação entre algo compreensível extensível a algo não conhecido" durante o processo de busca.

Esta organização possui alguns requisitos para sua construção como explica o autor, são eles: **Localização**, **Alfabeto**, **Tempo**, **Categoria** e **Hierarquia** dos dados Wurman (1997), com a finalidade de arquitetar, construir e moldar um ordenamento reconhecido como apropriado pela cognição humana. A **Localização** indica o lugar ou armazém onde os documentos com os dados necessários e desejados estão, por vezes, espalhados sem um ordenamento compreensível ao homem ou às máquinas; **o Alfabeto** possui uma equivalência estrutural com o dicionário de dados, a manifestação organizada e sistemática de palavras construídas de símbolos de um alfabeto finito; o **Tempo** indica o período inicial e final onde a pesquisa sobre os dados será restrita para delimitar um escopo de busca e de avaliação dos resultados; a **Categoria** manifesta a obtenção de termos e de grupos de termos a partir de significados equivalentes; a **Hierarquia** manifesta as relações taxonômicas ou não-taxonômicas e que podem ser tornar essenciais na busca de mais elementos facilitadores da ontologia.

Muitos autores, dentre eles Siqueira (2012) e Hessen (2003), reconhecem o fenômeno de construção do conhecimento a partir da correlação do Sujeito e Objeto onde a realidade (Objeto) é acessível a partir da experiência e do pensamento humano (Sujeito). Assim, antes de organizar, é preciso entender a informação a ser tratada nesses sistemas, restringir objetivamente o ambiente da busca, compreender o conjunto de palavras e

expressões no dicionário e/ou alfabeto que fazem parte do domínio, seus significados e relacionamentos para buscar os recursos inteligentes disponíveis para o projeto estrutural do ambiente de informacional. A aplicação dos requisitos de AI na Mineração de Texto e na aplicação de ferramentas de Aprendizado de Máquina permite que a observação sobre Objeto e Sujeito se torne mais acessível e precisa, garantindo otimização em relação a tempo e recurso, interoperabilidade e raciocínio dedutivo para a ontologia (GUIDALIA, *et al.*, 2023).

A expressão Inteligência Artificial, segundo alguns autores (NILSSON, 2009 e WANG, 2019), se refere à capacidade das máquinas de resolver problemas complexos, adaptando-se a um ambiente, mesmo com conhecimento e recursos limitados. Kaplan e Haenlein (2019) definem a IA como a capacidade que um sistema inteligente tem para interpretar corretamente dados, aprender com esses dados e usar o aprendizado para atingir metas e tarefas específicas por meio de uma adaptação flexível.

O Aprendizado de Máquina, uma área da IA, permite o aprendizado a partir de experiências passadas utilizando o princípio da inferência, a indução, para extrair conclusões genéricas a partir de um conjunto particular de exemplos (FACELI *et al*., 2021). Outro fator na utilização do AM é ser possível trabalhar com dados imperfeitos, como um conjunto de dados com a presença de ruídos, dados inconsistentes, ausentes e redundantes.

O desenvolvimento de um modelo no AM, isto é, regras de aprendizado, deve ser suficientemente robusto e genérico para que essas regras se apliquem e sejam válidas no conjunto de dados treinamento ou dados que estejam fora do conjunto treinamento, mas dentro do domínio ou contexto. As tarefas do AM podem ser divididas em **preditivas** ou **descritivas**.

As tarefas de predição incluem a capacidade de predizer o valor de um objeto a partir de um conjunto de atributos preditivos reconhecidos de um determinado treinamento; as tarefas de descrição extraem padrões dos valores preditivos de um conjunto de dados, procurando objetos similares entre si ou buscando regras de associações entre objetos do mesmo domínio (FACELI *et al*., 2021). Os autores destacam ainda que “um modelo preditivo pode gerar descrições de um conjunto de dados, e um descritivo pode prover previsões após ser validado”.

Algoritmos não supervisionados de AM, como o Apriori, são algoritmos descritivos de associação eficiente, responsáveis pela mineração de itens frequentes (*ItemSet*) para descoberta do conhecimento em regras de associação entre itens frequentes em base de dados que contém várias transações. É válido destacar que cada transação "suporta" um subconjunto específico de itens frequentes (AGRAWAL; SRIKANT,1994). Para os autores Katti Faceli et al. **(**2021), dar suporte é "testemunhar a favor" de um determinado conjunto de itens frente ao conjunto de dados.

Dessa forma, o Suporte de um *itemset* é a fração de transações condida por uma base de dados. No Apriori, considerando o conjunto de itens frequentes, é possível derivar regras de associação entre eles de natureza probabilística, na forma "se antecedente $então$ consequente". Para se chegar ao grau de incerteza ou certeza da regra, verifica-se a Confiança e o *Lift* (a sustentação) da regra (GORAYEB; DUQUE, 2024), que estabelece a descoberta de regras de associação por sua característica eficiente, de modo que cada uma delas suporte a associação entre um conjunto de itens frequentes (ALPAYDIN, 2014; SUMITHRA; PAUL, 2010).

No contexto de indução de modelos, cada algoritmo de AM representa possíveis viés de representação ou de busca, garantindo aprendizado e generalização do conhecimento adquirido no processo de treinamento e que serão utilizados no "Ambiente de Aprendizado de Ontologias". Na ontologia, a classificação nasce da necessidade de definir o objeto de busca, uma definição para cada termo de interesse, agrupando definições similares. Isso torna essa classificação em categorias de interesse, ganhando riqueza semântica e se relacionando numa estrutura hierárquica, agregada, com relações definidas.Aapresenta-se, assim, algo que possa ser compartilhado entre pessoas e sistemas, algo estruturado, preparado para produzir um sistema organizado de conhecimento.

Uma ontologia é construída para concentrar as regras de relacionamento entre os itens ou entre os termos chaves. Para Mori (2009), a ontologia relaciona conceitos representados formal e consensualmente dentro de um determinado domínio, é uma parcela de realidade reproduzida de maneira lógica com a qual podem operar diferentes sistemas de informação informatizados.

A conceitualização serve como vocabulário, é uma linguagem que permite que o recurso informacional seja utilizado e reutilizado, interoperável e aplicado por motores de busca de palavras-chaves para a recuperação da informação. Segundo Gruber (1995), em ontologias, os conceitos são formalizados por meio de classes de objetos contendo propriedades (ou atributos). Existem funções e relações na forma de um conjunto de asserções usadas para modelar um determinado domínio, definir vocabulário usado pela aplicação (ASTROVA; KOSCHEL; LEE, 2020) para propor axiomas sobre estes elementos como meio de compartilhar o conhecimento (EVARISTO; DUQUE, 2011).

Formalmente, a ontologia pode ser descrita como a seguinte tupla:

$$O = < C, H, R, A >$$

Onde O representa ontologia, C representa um conjunto de classes (conceitos), H representa um conjunto de ligações hierárquicas entre os conceitos (relações taxonômicas), R representa um conjunto de vínculos conceituais (relações não taxonômicas) e A representa um conjunto de regras e axiomas (ZOUAQ; GASEVIC; HATALA, 2011).

Alguns estudos do tema engenharia de ontologia defendem um processo de construção correspondente a um ciclo de vida, compartilhado por várias metodologias difundidas em CI, e inclui alguns estágios distintos (FOX et al., 1993), (NOY; MCGUINNESS, 2001), (FERNÁNDEZ; GÓMEZ-PÉREZ; JURISTO, 1997) e (USCHOD; GRUNINGER, 1996) para extração e para classificação dos elementos relevantes, da implementação, do compartilhamento e da expansão cíclica e incremental do conhecimento, contanto com as seguintes etapas:

Quadro 1: Etapas de construção de ontologias

Definição do escopo, requisitos e conhecimento específico	São definidos o plano de projeto e as questões de competência para o modelo de ontologia como meios de acompanhamento dos objetivos do projeto;
Modelagem	Encadeamento de conceitos e fluxo informacional, regras, padrões de qualidade, restrições e outros mecanismos do processo de

	classificação que apoiam a representação da realidade;
Formalização e implementação	utilização da linguagem lógica para representar a conceitualização;
Validação	estabelecimento dos critérios para relevância dos conceitos e apresentação do conteúdo informacional como a generalidade do modelo, a eficiência e a clareza no raciocínio, na trasformabilidade, na escalabilidade e na integração do modelo para as práticas do conhecimento;
Evolução	avaliação do grau de manutenibilidade e atualização do modelo buscando expansão do conhecimento.

A incorporação do Aprendizado de Máquina no processo de construção da ontologia, como *Ontology Learning,* busca extração do conhecimento novo, útil e relevante de um conjunto de dados e trata a informação desde o início do processo de construção. O AM permitirá utilizar meios descritivos de algoritmos de AM para definir conceitos, hierarquias e regras, e no final os meios preditivos do AM permitirão predizer descrições e sentenças úteis do objeto de interesse do domínio.

Dessa forma, a *Ontology Learning* relaciona descoberta de regras, enriquecimento, aprendizagem aprimorada, proposta automática da taxonomia e de relações não taxonômicas (HASSAN; RASHID, 2021). Segundo os autores, técnicas assistidas por IA podem oferecer a classificação de padrões e mineração de conhecimento descobrindo vários tipos de relações ocultas no conhecimento, como por exemplo, a relação "**termo principal-termos atributos"** do objeto de interesse, relação "**termos-comportamento"** do objeto de interesse e a relação entre "**termo principal-termos complementares"** não taxonômicos, mas ainda assim, dentro do domínio onde o objeto se encontra.

A extração e a organização de conceitos e o conhecimento significativo são fundamentais para a compreensão da máquina e para a capacidade de

raciocínio de algoritmos, já que a *Ontology Learning* se encarrega da extração, representação e refinamento de ontologias estruturadas que encapsulam as complexidades de vários domínios (DU *et al*., 2024). Ainda segundo os autores, técnicas de aprendizado possibilitam descrever termos de um determinado domínio a partir da compreensão semântica e inferir relações entre entidades tal qual a proposta descritiva e preditiva de AM.

Ontology Learning refere-se ao suporte semiautomático ou automático para a construção, instanciação e evolução de uma ontologia (HASSAN; RASHID, 2021) através da descoberta de conhecimento em diferentes tipos de fontes de dados e com sua representação por meio de uma abordagem inteligente para automatizar ou semi-automatizar o processo a partir de dados brutos (documentos de texto, imagens, dados numéricos ou mesmo outras ontologias).

A Descoberta de Conhecimento em Bancos de Dados ou *Knowledge Discovery in Databases* (KDD) é o processo de identificação de novos conhecimentos extraídos de um banco de dados ou de textos na Mineração de Texto (MT) (Fayyad et al., 1996). O processo de construção de ontologias, *Ontology Learning,* consiste em um conjunto de camadas com conceitos, relações e axiomas extraídos de texto não estruturado: mineração do termo, definição de significado, conceito e hierarquia, associação das relações, extração do conhecimento para axiomas e regras, como evidenciado na Figura 1:

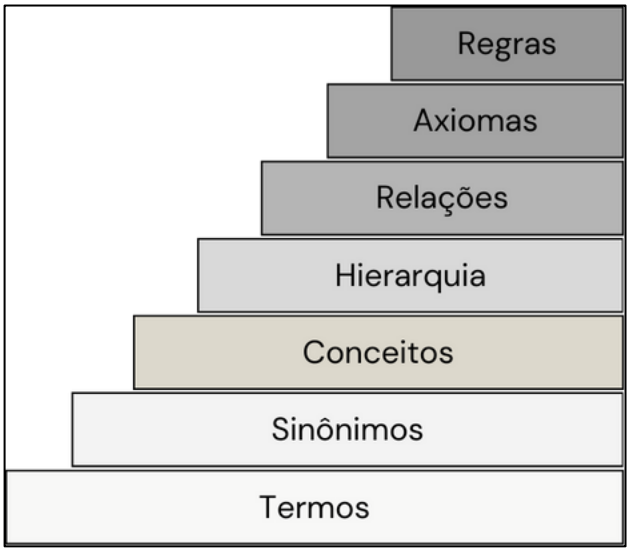


Figure 1. Modelo de *Ontology Learning Layer Cake*, Fonte: Adaptado de (BUITELAAR; CIMIANO; MAGNINI, 2005).

A camada de **Termos**, segundo os autores, seleciona termos relevantes ao domínio por definição em documentos ou entrevistas com especialistas ou por meio do processo estatístico em mineração de texto. A camada de **Sinônimos** aborda a aquisição de variantes semânticas de termos, ou seja, termos que são semelhantes em significado, ou recíproco do objeto de interesse. A camada de **Conceito** parte da designação, da indução ou da formação da intenção de algo e descreve a realização de todas as instâncias se sobrepondo aos termos e sinônimos (LISI, 2007), o que envolve organizar termos relacionados em hierarquias ou categorias com base em suas similaridades, funcionalidades ou relações semânticas (DU*, et al.*, 2024). A camada de **Hierarquia** de conceitos trata das relações entre palavras que têm significados mais específicos e outras com significados mais gerais (hiponímia), garantindo as futuras relações léxico-sintáticas para que as descrições do objeto de interesse em determinado domínio façam sentido e estejam bem estruturadas (BUITELAAR; CIMIANO; MAGNINI, 2005). A etapa de **Relação** (não hierárquica) depende da Mineração de Termos e relaciona os termos essenciais com outros tantos complementares, buscando aprimorar o sentido léxico-sintático para a descrição do objeto, criando camadas de novos conceitos e hierarquias, obtidas por meio de regras de associação (MAEDCHE; STAAB, 2000). As etapas de **Axiomas ou Regras** estão relacionadas ao trabalho de restringir as relações por meio de conhecimento especializado ou dos algoritmos de AM para extrair as regras e identificar *Data Properties* durante o processo de Mineração de Texto (GORAYEB; DUQUE, 2024). Elas definem dependências ou relações lógicas entre entidades ou conceitos, buscando formalizar o conhecimento do domínio e estabelecer restrições lógicas dentro da ontologia.

3. Trabalhos relacionados

Combinações de Ontologia e Aprendizado de Máquina e ambientes de *Ontology Learning* apresentam muitas categorias de estudo: mapeamento de ontologias relacionadas; enriquecimento de termos da ontologia; população

automática de ontologias; descoberta automática de regras da ontologia; construção automática de relações taxonômicas e não taxonômicas; e redução da granularidade dos conceitos de ontologia com aprimoramento do raciocínio.

Nestas categorias de estudo existem basicamente duas formas de aprendizado para a ontologia: semiautomática e automática e muitos desafios para analisar como: intensidade do trabalho; formulação de axiomas; aquisição automática do conhecimento; escalabilidade; heterogeneidade; avaliação e validação; e ambiguidades (DU*, et al*., 2024) e para Guidalia, ***et al***. (2023): diversidade dos métodos de AM aplicados; raciocínio indutivo e dedutivo; e construção automática da taxonomia.

Levando em consideração a utilização de algoritmos de AM não supervisionados para descoberta de relações e de estatísticas na obtenção dos dados de construção, população e enriquecimento das ontologias, os artigos abaixo foram avaliados para análise da Arquitetura da Informação proposta na organização dos dados de entrada para a *Ontology Learning*, Quadro 2 abaixo:

Quadro 2: Avaliação da Arquitetura da Informação (AI) em estudos para *Ontology Learning* usando algoritmos de AM não supervisionado

Autores	**Arquitetura da Informação (AI) para aprendizado da ontologia**	**Aplicação**	**Resultados da AI**
Chung; Yoo; Choe, (2020)	Arquitetura baseada em *data mining process*: 1. Compressão dos dados; 2. Seleção dos dados; 3. Preparação dos dados; 4. Modelagem dos dados; 5. Avaliação dos dados.	Avaliação de risco na área de saúde da pessoa.	Os dados são tratados somente do ponto de vista computacional durante o processo de Mineração de Dados e Mineração de Texto.
Yang, (2020)	Arquitetura baseada na organização racional: raciocínio, análise semântica dos termos e utilização de especialistas para extração dos conceitos (etapas da engenharia de ontologia).	Avaliação do risco para logística na segurança financeira.	Não há uma preparação dos dados antes da engenharia de ontologia.

Maedche; Staab, (2000)	Arquitetura baseada em um componente de gerenciamento de texto e processamento “para selecionar textos de domínio explorados no processo de descoberta” e escolha de métodos de pré-processamento de texto.	Avaliação de técnicas de AM utilizando aprendizado superficial e profundo.	Não há uma preparação dos dados antes do pré-processamento de texto para a entrada no ambiente de modelagem da ontologia - *Ontology Modeling Environment* (OntoEdit4).
HARIDY, *et al.*, (2023)	Arquitetura baseada na engenharia de ontologia com a primeira fase 1. Requisitos de aquisição de dados: a. Identificação; b. Análise; c. Especificação.	Adquirir uma lista de requisitos que a ontologia de saída com a identificação das principais informações do domínio e das necessidades do usuário na área de Turismo.	Os dados são formatados como requisitos da informação na primeira etapa da engenharia de ontologia (ON-ODM *methodology*).

Os trabalhos sempre destacam a importância da qualidade dos dados do Corpus para construção das ontologias de domínio semiautomáticas ou automáticas e como a baixa qualidade dos dados influencia na falta de aprofundamento das técnicas de PLN para extração e integração dos dados e das relações sintáticas entre palavras dificultando o aprofundamento e expressividade no domínio e evidências sobre a relevância de conceitos. Apesar disto, as metodologias estudadas não se concentram em fases de conhecimento ou organização da informação, características da AI, gerando um número grande de termos não adequados como elementos da ontologia de domínio.

4. Desenvolvimento do Modelo de Arquitetura da Informação para *Ontology Learning*

O modelo de Arquitetura da Informação proposto na Figura 1 independe de uma ferramenta ou *framework* específico. O processo pode ser aplicado a qualquer conjunto de dados de um determinado domínio e nessa perspectiva genérica, a entrada passa por requisitos da AI para conhecer e organizar um Corpus dedicado à extração de conceitos relevantes à um determinado domínio existente. A saída é uma lista de conceitos e relacionamentos de interesse da ontologia.

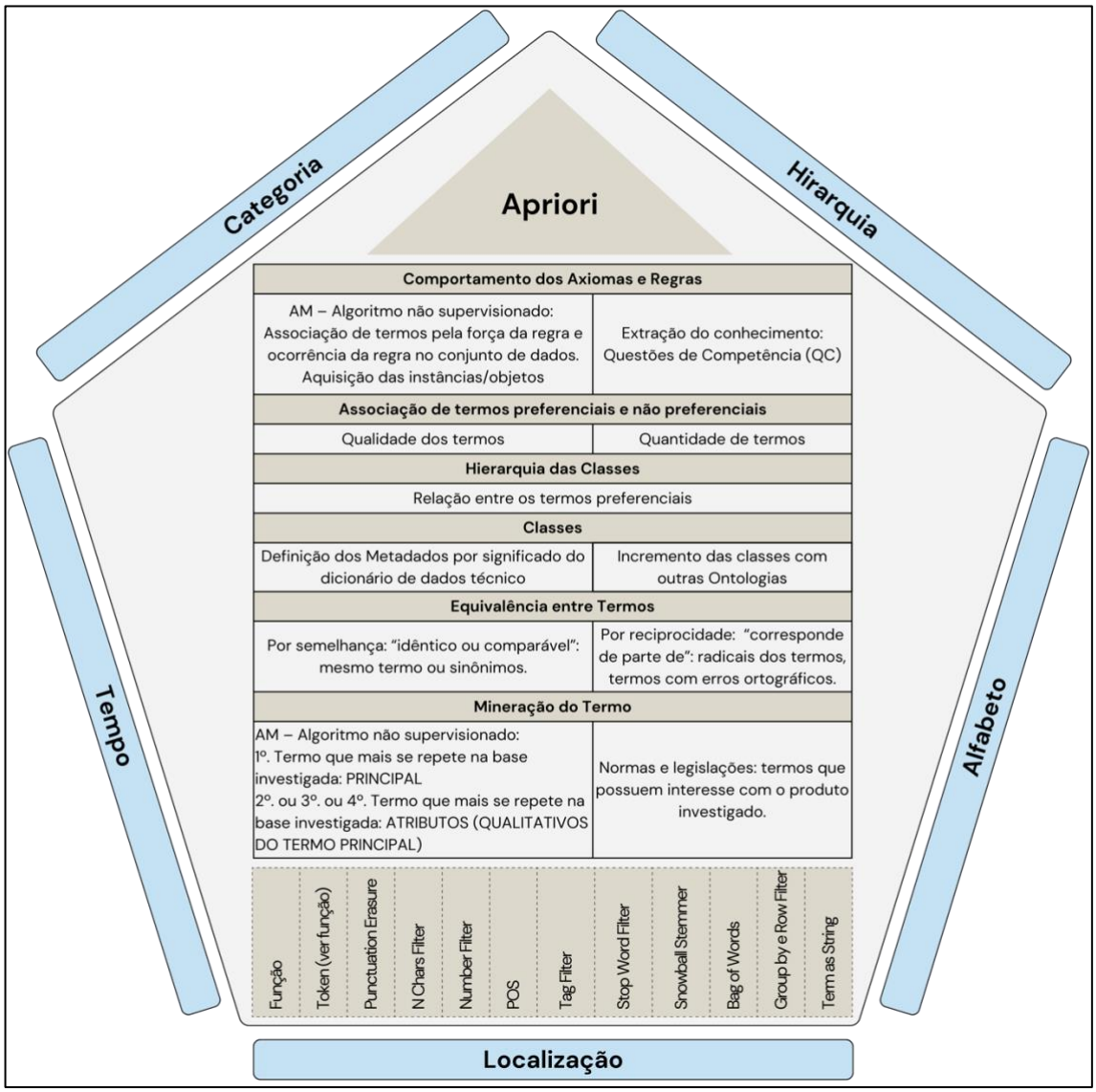


Figura 2: Modelo de Arquitetura da Informação para *Ontology Learning*.

A proposta de um modelo de AI para organizar dados nas camadas de construção da *Ontology Learning* alcança as fases de Mineração de Textos e de aplicação de técnicas de AM como PLN e algoritmos não supervisionados para estatísticas e regras de associação e serve como um manual de "boas práticas"

para identificar quais “recursos de entrada” são relevantes, em qual formato estão, como devem ser manuseados e onde se encontram, isto é, a disponibilidade do conhecimento prévio na entrada da construção da ontologia.

O conhecimento prévio usado na camada ‘Mineração do Termo’, por exemplo, pode ser transformado **na primeira versão da primeira camada da *Ontology Learning*** (Figura 3) apresentando uma estatística sobre termos minerados que posteriormente serão classificados estatisticamente em termo principal, atributos e complementares.

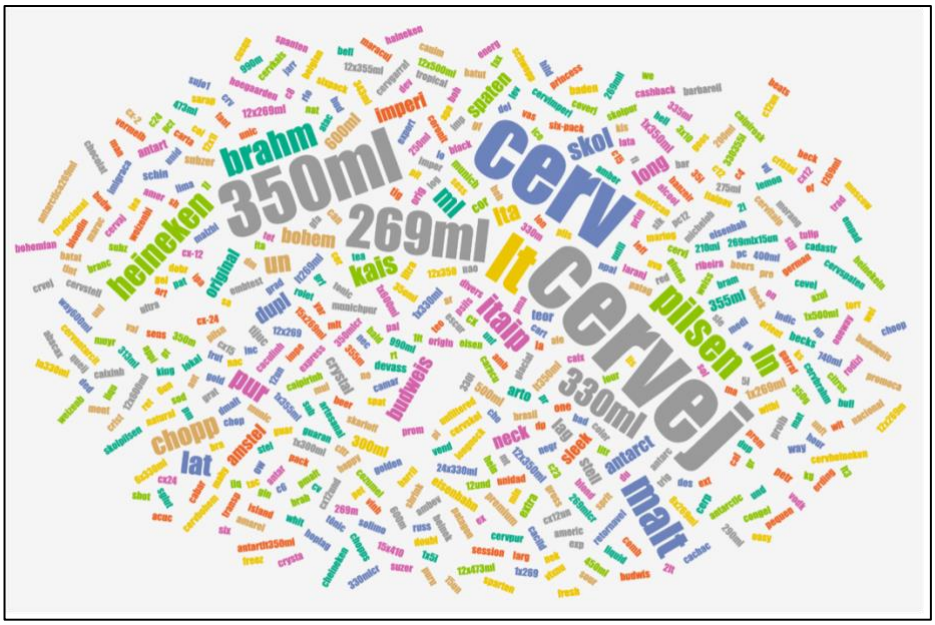


Figura 3: Camada ‘Mineração do Termo’, nuvem de termos minerados a partir da AI na amostra do produto de interesse cerveja.

A versão de saída da primeira camada ‘Mineração do Termo’, já classificada, será submetida a um novo ciclo de AI (Localização, Alfabeto, Tempo, Categoria, Hierarquia) para a entrada da próxima camada ‘Equivalência entre Termos’ e assim sucessivamente para todas as demais camadas de construção da *Ontology Learning*, favorecendo o aprendizado em um processo iterativo e incremental (Figura 1). Ao final do percurso, o construtor tem a liberdade de reiniciá-lo e reproduzi-lo quantas vezes for necessário para acrescentar fontes de dados e informações e extrair novos termos, conceitos, hierarquias, axiomas e regras para a ontologia.

Considerando o ambiente de NFC-*e,* o desafio é minerar termos, no campo da nota fiscal de **descrição do produto,** um campo livre e sem estruturas lógicas computacionais e sem metadados, que representem o produto de interesse: cerveja, de tal forma que seja possível relacionar o termo principal cerveja ao demais termos que o qualificam em quantidade e qualidade suficientes para reconhecer formas úteis de descrição do produto ao final da ontologia. Além da descrição útil do produto comercializado, é necessário concatenar na ontologia outros dados lógicos, computacionalmente estruturados, que representem informações de venda como valor, unidade, emitente da nota, consumidor etc. construindo uma ontologia adequada para o acompanhamento da comercialização da mercadoria pelos órgãos competentes estaduais.

Os recursos de entrada do domínio de NFC-*e* são de três tipos: estruturados, semiestruturados e não-estruturados:

a. Os recursos estruturados estão localizados em esquemas de banco de dados da NFC-*e* (de onde foi tirada a amostra de dados do estudo) ou outras ontologias existentes de venda de produtos e de cerveja (fonte de pesquisa). A principal questão no aprendizado de *Ontology Learning* com fontes de informação estruturadas é determinar quais peças de informação estruturada podem fornecer conhecimento adequado e relevante. Utilizando o modelo proposto da AI o conhecimento selecionado para os recursos estruturados são: linhas 1, 2, 3 (item 2 do resultado selecionado) e 4 do Quadro 3;
b. Os recursos semiestruturados são tipicamente representados por dicionários eletrônicos compilados manualmente, geralmente são fontes abertas, bem documentadas e gratuitas. Utilizando o modelo proposto da AI o conhecimento selecionado para o recurso semiestruturados é: linha 3 do Quadro 3;
c. Os recursos não estruturados dependem de técnicas de AM, processamento de linguagem natural e recuperação de informações MT para buscar padrões e tendências relevantes ao domínio. O resultado para esta pesquisa, segundo a AI, estão nas linhas 5 e 6 do Quadro 4.

Quadro 3: Ciclo na primeira camada ‘Mineração do Termo’ do Modelo de Arquitetura da Informação (AI) da *Ontology Learning* para produto cerveja NFC-*e*.

#	Tipos de dados	Requisito de AI	Resultado selecionado
1	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para a identificação do produto:	**Localização**	1. NCM; 2. cEAN (código GTIN); 3. Número do item; 4. Quantidade; 5. Unidade; 6. Valor; 7. Descrição original (extraído do campo “descrição do produto” da nota fiscal).
2	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para identificação da venda:		8. Unidade da Federação (UF); 9. Município; 10. Emitente (Grupo C da NF-*e*); 11. Destinatário (Grupo E da NF-*e*); 12. Número da Nota Fiscal ou Cupom Fiscal.
3	Gramática com termos (léxico) que descreve o produto cerveja e que são inicialmente identificados.	**Alfabeto**	1. Extraídos da Resolução Estadual n.° 0028/2023; 2. Extraídos das ontologias de referência.
4	Período de interesse para auditoria das NFC-*e*.	**Tempo**	1. Período da amostra entregue pela SEFAZ/AM: 01/02/2023 a 31/05/2023.
5	Campos não estruturados da “descrição do produto”: classificação por equivalência ou reciprocidade de palavras relacionadas aos termos que descrevem a cerveja (pela estatística).	**Categoria**	1. Termo de interesse: CERVEJA.
6	Campos não estruturados da “descrição do produto”: associações construídas entre os termos do alfabeto (pela	**Hierarquia**	*Null*

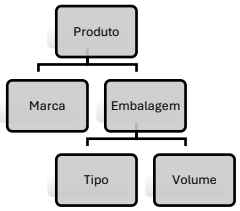
	força das regras entre termos).		

Ao final do ciclo para a primeira camada “Termo”, o requisito da AI **Hierarquia** não foi utilizado, pois não há conhecimentos suficientes sobre os dados para definir quais hierarquias serão úteis e que devem ser mineradas para a construção da ontologia. Além disso, o modelo de AI orienta que a única categoria que faz parte do domínio e que é conhecida é o próprio objeto de interesse: CERVEJA.

Ao aplicar o modelo de AI para outras camadas da *Ontology Learning*, Quadro 4, os requisitos começam a se completar e apresentar orientação para organizar a informação e recuperá-la na forma necessária à construção da ontologia de NFC-*e*.

Quadro 4: Ciclo completo do Modelo de Arquitetura da Informação (AI) da *Ontology Learning* para produto cerveja NFC-*e*

#	Tipos de dados	Requisito de AI	Resultado selecionado
1	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para a identificação do produto:	**Localização**	13. NCM; 14. cEAN (código GTIN); 15. Número do item; 16. Quantidade; 17. Unidade; 18. Valor; 19. Descrição original (extraído do campo “descrição do produto” da nota fiscal).
2	Campos lógicos e computacionalmente reconhecidos em linhas e colunas das tabelas do banco de dados da NFC-*e* para identificação da venda:		20. Unidade da Federação (UF); 21. Município; 22. Emitente (Grupo C da NF-*e*); 23. Destinatário (Grupo E da NF-*e*); 24. Número da Nota Fiscal ou Cupom Fiscal.
3	Gramática com termos (léxico) que descreve o	**Alfabeto**	3. Extraídos da Resolução Estadual n.° 0028/2023; 4. Extraídos das ontologias de referência.

	produto cerveja e que são inicialmente identificados.		
4	Período de interesse para auditoria das NFC-*e*.	**Tempo**	2. Período da amostra entregue pela SEFAZ/AM: 01/02/2023 a 31/05/2023.
5	Campos não estruturados da "descrição do produto": classificação por equivalência ou reciprocidade de palavras relacionadas aos termos que descrevem a cerveja (pela estatística).	**Categoria**	2. Exemplo identificado: CERVEJA, CERVEJ, CERV; LATA, LTA, LT; LONGNECK, LONG, LN.
6	Campos não estruturados da "descrição do produto": associações construídas entre os termos do alfabeto (pela força das regras entre termos).	**Hierarquia**	1. Exemplo identificado: "CERVEJA + SKOL + LATA + 269ML" " "+ "SKOL + LONGNECK + 350ML". 2. Exemplo de Metadados: 

A aplicação da AI permite uma análise semântica profunda extraindo, portanto, vários blocos de conhecimento como: termos de domínio, relações do tipo "é um" e relações conceituais que são então validados e exportados para uma ontologia de domínio

5. Resultados e discussões

O objetivo desta sessão é apresentar e discutir a avaliação dos resultados da Arquitetura da Informação utilizada para organizar dados e que serão recuperados por Mineração de Texto e utilizados por Aprendizado de Máquina na construção de uma ontologia sobre o produto cerveja.

O resultado, baseado nos requisitos propostos ao final da primeira camada 'Mineração do Termo' para *Ontology Learning* (Quadro 3), bem como ao final de todo ciclo (Quadro 4) representa uma resposta de cada elemento da AI como

uma orientação para construção da ontologia, sem a necessária presença do especialista do domínio. Esta sobreposição de papéis, organização da informação da AI por meio dos elementos formadores de uma arquitetura concreta sobre a subjetividade do especialista poderá garantir eficiência e economia de tempo na recuperação dos dados durante o processo de mineração e análise do texto com suas relações.

Além disso, dos trabalhos relacionados no Quadro 2, é possível analisar que eles não apresentam um modelo de Arquitetura da Informação anterior à construção da ontologia (etapas do Quadro 1) para auxiliar no processo de organização, recuperação e definição dos termos durante a mineração. As técnicas utilizadas são de pré-processamento de mineração de dados e possuem a finalidade de limpar os dados já coletados aleatoriamente das bases exploradas. Alguns trabalhos apresentam como fase inicial o levantamento de requisitos, mas não há um modelo de orientação e boas práticas para a construção de requisitos funcionais de busca de termos e relações.

Por fim, aplicar a AI de forma iterativa e incremental aumentará a possibilidade de inserir novos termos e conceitos na ontologia, deixando na documentação dos requisitos o caminho utilizado para definição de classes e relações adicionais. Com isso, evita-se informação desnecessária ou duplicada.

6. Considerações finais

Ontologias são instrumentos organizados de conhecimento para domínios e interesses diversos. Para superar tempo e dificuldades do processo de desenvolvimento, o Aprendizado de Máquina é amplamente utilizado, incorporando técnicas ao processo de Mineração de Dados e de Texto para a construção da *Ontology Learning.* Contudo, há um grande desafio de escolher e organizar quais dados e informações são significativas para acelerar o aprendizado e tornar o processo eficiente.

A entrada de dados para desenvolvimento de ontologias em geral é apresentada por especialistas do domínio, em *Ontology Learning,* cuja entrada é também derivada da Mineração de Dados e Mineração de Texto feita sobre o volume de dados estruturados, semiestruturados e não estruturados. No entanto, ainda há

certas atividades de organização dos dados e da informação que são essenciais para a recuperação de termos, conceitos ou relações hierárquicas e que não foram cobertas nos modelos de construção de ontologia avaliados durante a pesquisa. Com vista a isso, em outros casos, as atividades poderiam ser aprimoradas.

O Modelo de AI para *Ontology Learning* concentra-se na forma de aquisição de requisitos, que contribui significativamente para o resultado. Os requisitos arquiteturais são reunidos sob diferentes perspectivas e apresentados em 5 formas diferentes, quais sejam: Localização, Alfabeto, Tempo, Categoria e Hierarquia. O Modelo é iterativo e incremental para servir de referência, apontando dados e informações para uso da Mineração de Texto e algoritmos inteligentes, para que o aprendizado seja mais eficiente, uma vez que o conjunto de termos e suas associações ficarão mais visíveis a cada interação.

A melhora da modelagem conceitual, ao incorporar as teorias da Arquitetura da Informação, para construir um artefato de engenharia de requisitos, é a etapa principal no ciclo de construção de ontologias, inferindo caminhos preferenciais para as técnicas de mineração e PNL e extraindo uma lista de novos candidatos de termos e relações a cada finalização do ciclo.

Por fim, uma aplicação para o ambiente de notas fiscais eletrônicas para o produto cerveja foi apresentada, gerando os resultados e facilitando o entendimento do processo.

REFERÊNCIAS

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules *In*: STONEBRAKER, M.; HELLERSTEIN, J. M. (ed.) **Readings in database systems**. 3. ed. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1994. p. 580–592. Disponível em: https://www.vldb.org/conf/1994/P487.PDF. Acesso em: 13 jan. 2025.

ALPAYDIN, E. **Introduction to Machine Learning**. 3 ed. Massachusetts: MIT Press, 2014. Disponível em: https://dl.matlabyar.com/siavash/ML/Book/Ethem%20Alpaydin-Introduction%20to%20Machine%20Learning-The%20MIT%20Press%20(2014).pdf. Acesso em: 13 jan. 2025.

ASTROVA, I.; ARNE KOSCHEL, A.; LEE, S. How the Apriori Algorithm Can Help to Find Semantic Duplicates in Ontology. ***Joint Conference on Knowledge-Based Software Engineering***. [*S. l.*]: JCKBSE, 2020). Disponível am: https://www.researchgate.net/publication/343195533_How_the_Apriori_Algorithm_Can_Help_to_Find_Semantic_Duplicates_in_Ontology. Acesso em: 13 jan. 2025.

BABAEI GIGLOU, H.; D'SOUZA, J.; AUER, S. LLMs4OL: Large language models for ontology learning. In: **International Semantic Web Conference**. Cham: Springer Nature Switzerland, 2023. p. 408-427.

BUITELAAR, P.; CIMIANO, P.; MAGNINI, B. Ontology learning from text: An overview. **Ontol. Learn. from text Methods***,* Amisterdã, v. 123, p. 3–12, 2005. Disponível em: https://pub.uni-bielefeld.de/record/2497696. Acesso em: 7 jul. 2024.

CHUNG, K.; YOO, H.; CHOE, D. Ambient context-based modeling for health risk assessment using deep neural network. **Journal of Ambient Intelligence and Humanized Computing**, Germany, v. 11, p. 1387–1395, 2020. DOI 10.1007/s12652-018-1033-7. Disponível em: https://www.researchgate.net/publication/327650836_Ambient_context-based_modeling_for_health_risk_assessment_using_deep_neural_network. Acesso em: 13 jan. 2025.

DU, R.; AN. H.; WANG, K.; LIU, W. A short review for ontology learning: Stride to large language models trend. **arXiv preprint arXiv:2404.14991**, Ithaca, Nova York, v. 1, p. 11-24, 2024. Disponível em: https://arxiv.org/abs/2404.14991. Acesso em: 13 jan. 2025.

EVARISTO, E.; DUQUE, C. G. Recuperação da Informação em vídeo por meio de análise multimodal. *In*: DUQUE, C. G. (org). **Ciência da Informação Estudos e Práticas**. Brasília, DF: Thesaurus Editora de Brasília Ltda, 2011. p. 237-250.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. de L. F. de. **Inteligência artificial:** uma abordagem de aprendizado de máquina. Rio de Janeiro, Editora LTC, 2021.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. Da Mineração de Dados à Descoberta de Conhecimento em Bancos de Dados. **AI Magazine**, Arlington, USA, v. 17, n. 3, p. 37-42, 1996. DOI 10.1609/aimag.v17i3.1230. Acesso em: 12 ago. 2024.

FERNÁNDEZ, M.; GÓMEZ-PÉREZ, A.; JURISTO, N. METHONTOLOGY: From Ontological Art Towards Ontological Engineering. **Proceedings of the Ontological Engineering AAAI97**. Palo Alto, EUA: [*S.n.*], 1997. P. 25-43. Disponível em: https://aaai.org/papers/0005-ss97-06-005-methontology-from-

ontological-art-towards-ontological-engineering/. Acesso em: 27 set. 2023.

FOX, M. S**.** The TOVE Project Towards a Common-Sense Model of the Enterprise. *In:* BELLI, F.; RADERMACHER, F. J. (ed.) **Industrial and engineering applications of artificial intelligence and expert systems**. London: [*S.n.*], 1993. p. 25–34.

GHIDALIA, S. *et al.* Combining machine learning and ontology: A systematic literature review. **arXiv preprint arXiv:2401.07744**, Ithaca, Nova York, v. 2, 2024. p. 47-55. Disponível em: https://arxiv.org/abs/2401.07744. Acesso em: 2 set. 2024.

GORAYEB, D. M. da C.; DUQUE, C. G. Proposta de metadados para descrição de produtos da Nota Fiscal de Consumidor Eletrônica (NFC-e) usando Apriori. **P2P & Inovação**, Rio de Janeiro, v. 11, n. 1, p. 1-24, e-7124, jul./dez.2024. DOI 0.21728/p2p.2024v11n1e-7124. Acesso em: 13 dez. 2024.

GRUBER, T. R. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal Human-Computer Studies**, Padova, Italy, v. 43, 1995, p. 907-928. DOI 10.1006/ijhc.1995.1081. Acesso em: 2 set. 2024.

GRÜNINGER, M.; FOX, M. S. The Role of Competency Questions in Enterprise Engineering. *In:* ROLSTADÅS, A. (ed.). **Benchmarking**: Theory and Practice. Boston: IFIP Advances in Information and Communication Technology, 1995. p. 22-31.

HAENLEIN, M., KAPLAN, A. A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. **California Management Review**, Califronia, n. 61, p. 5-14, 2019. DOI 10.1177/0008125619864925. Acesso em: 12 ago. 2024.

HARIDY, S.; ISMAIL, R. M; BADR, N.; HASHEM, M. An Ontology Development Methodology Based on Ontology-Driven Conceptual Modeling and Natural Language Processing: Tourism Case Study. **Big Data and Cognitive Computing**,Basel Suíça, ano 7, v. 101, n. 2, p. 1-23, 2023. DOI 10.3390/bdcc7020101. Acesso em: 7 maio 2023.

HASSAN, B. A.; RASHID, T. A. Artificial intelligence algorithms for natural language processing and the semantic web ontology learning. **arXiv preprint arXiv:2108.13772**, Ithaca, Nova York, p. 22-41, 2021. Disponível em : https://arxiv.org/abs/2108.13772. Acesso em: 2 ago 2024.

HESSEN, J. **Teoria do conhecimento**. Tradução João Vergílio Gallerani Cuter; revisão técnica Sérgio Sérvulo da Cunha. 2. ed. São Paulo: Martins Fontes, 2003.

KUROKI JUNIOR, G. H.; DUQUE, D. G. Arquitetura da informação aplicada ao processamento de linguagem natural: uma proposta.

Contribuições da Ciência da Informação no pré-processamento de dados para treinamento e aprendizagem de redes neurais artificiais. **RDBCI**, Campinas, SP, e.21, e023002, 2023. DOI 10.20396/rdbci.v21i00.8671396/30919. Acesso em: 14 jul. 2023.

LISI, F. A. Building Rules on Top of Ontologies for the Semantic Web with Inductive Logic Programming. **arXiv preprint arXiv:2108.13772**, Ithaca, Nova York, v. 8, n. 3, p. 47-60, 2007. https://arxiv.org/abs/0711.1814. Acesso em: 15 maio 2024.

MAEDCHE, A.; STAAB, S. Discovering Conceptual Relations from Text. *In:* **Proceedings of the 14th European Conference on Artificial Intelligence**. Horn, EUA: Ed. IOS Press, 2000. p. 321–325.

MORI, A. **Modelagem do conhecimento baseada em ontologias aplicada às Políticas Públicas e Habitação.** 2009. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação, Universidade de Brasília, Brasília, 2009. Disponível em: https://repositorio2.unb.br/jspui/handle/10482/6880. Acesso em: 14 jul. 2023.

NILSSON, N. J. **Inteligência Artificial**: uma abordagem moderna. 3. ed. Upper Saddle River, NJ: Prentice Hall, 2009.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101:** A guide to creating your first Ontology. California: Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Stanford Medical Informatics Technical Report SMI-2001-0880, 2001. Disponível em: https://www.researchgate.net/publication/243772462_Ontology_Development_101_A_Guide_to_Creating_Your_First_Ontology. Acesso em: 2 maio 2023.

SIQUEIRA, A. H. de. **Arquitetura da Informação**: uma proposta para a fundamentação e caracterização de uma disciplina científica. 2012. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, 2012.

SUMITHRA, R.; PAUL, S. **Using distributed apriori association rule and classical apriori mining algorithms for grid-based knowledge discovery**. **IEEExplorer**, Kadur, Índia, 2010, India. Disponível em: https://ieeexplore.ieee.org/document/5591577. Acesso em: 20 mar 2023.

USCHOLD, M.; GRUNINGER, M.Ontologies: principles, methods and applications. **Knowledge Engineering Review**, Nova York, v. 11, n. 2, p. 93-155, 1996. DOI 10.1017/S0269888900007797. Acesso em: 14 jul. 2023.

WANG, P. On defining artificial intelligence. **Journal of Artificial General Intelligence**, Boston, v. 10, n. 2, p. 1-37, 2019. DOI 10.2478/jagi-2019-0002. Acesso em: 7 jul. 2024.

WURMAN, R. S. **Information Architects**. [*S.l.*]: Graphis Inc, 1997.

YANG, B. Construction of logistics financial security risk ontology model based on risk association and machine learning. **Safety Science**, Califórnia, v. 123, 2020. DOI 10.1016/J.SSCI.2019.08.005. Acesso em: 20 mar 2023.

ZOUAQ, A.; GASEVIC, D.; HATALA, M. Towards open ontology learning and filtering. **Information Systems**, Illinois, EUA, n. 36, v. 7, p. 1064–1081, 2011. DOI 10.1016/j.is.2011.03.005. Disponível am: https://www.sciencedirect.com/science/article/abs/pii/S0306437911000391. Acesso em: 13 jan. 2025.