1 PRÉ-PROJETO DE PESQUISA

Laminar: sistema de mineração, organização e visualização de informação

1.1 Introdução

"Com a crescente popularização e disseminação da Web, um grande volume de dados tornou-se universalmente acessível para um número cada vez maior de usuários" (ARANTES, 2001), e com esse grande volume é esperado que parte das informações relevantes sejam difíceis de se encontrar. A proposta de pesquisa visa desenvolver um sistema que possa auxiliar no processo de extração, transformação e carregamento dos dados dispersos na internet, aprimorando a organização e visualização das informações extraídas de diferentes fontes. O sistema proposto se mostra relevante no passo em que visa explorar um campo de estudo que dispõe de profundas complexidades em cada uma de seus processos. Fazendo necessário compreender como lidar com a variabilidade de informações e suas disposições nos diferentes sites, gerir o acesso aos websites e evitar possíveis bloqueios pelo sistema, utilizar estruturas de dados adequadas para armazenar as informações extraídas, e apresentar essas informações de forma clara e objetiva para o usuário.

Este pré-projeto propõe a pesquisa nas áreas da teoria da informação; web scraping; mineração, visualização e ciência de dados; bem como inteligência artificial para processamento de linguagem natural. O objetivo é desenvolver um sistema de mineração de dados guiada por inteligência artificial para organização e visualização de informação, seguido do desenvolvimento de um sistema *web* para a aplicação prática do sistema proposto.

Seus objetivos incluem o desenvolvimento de um sistema para *web scraping*, processamento e visualização de informações estruturadas que permitam ao usuário a manipulação de dados relevantes e previamente esparsos, utilizando de tabelas, filtros, ordenações e outros métodos de manipulação de dados, para fim de encontrar as informações que deseja.

1.2 Referencial Teórico

Desde o início do século, diversos estudos vêm atacando o problema do grande volume de dados dispersos na internet (ARANTES, 2001; SILVA, 2002; CORREA, 2008). O processo de extração, transformação e carregamento dos dados, conhecido como ETL (*Extract, Transform, Load*), apresenta diversos desafios a ser superados, cada um com uma complexidade própria.

Em todas as etapas desse processo, é necessário considerar a variabilidade de informações e quão relevantes elas são para o que se deseja extrair. Esse conceito, abordado por Fortes (2022), aponta como exemplos de critérios de qualidade a acurácia, a novidade e a diversidade dos dados.

Diversos websites disponibilizam informações de forma dinâmica de acordo com uma estrutura baseada em HTML. Essa questão segue em direção dos trabalhos de **??**) e VENEROSO (2019). O primeiro, visa a extração de dados semi-estruturados de forma semi-automática, utilizando de gramáticas tabulares e de exemplos de dados como base para a extração. O segundo, propõe uma forma de compreender os dados dispostos em HTML, pois cita que a forma como os códigos HTML e CSS estão dispostos podem trazer informações quanto aos dados coletados.

Após a coleta dos dados brutos, já no escopo da transformação, ocorre o processamento desses dados. CARVALHO (2009) aborda alguns desafios presentes nessa etapa, sendo ele a integração dos dados obtidos de diferentes fontes. Ele utiliza de algoritmos para lidar com a deduplicação dos registros e também para a integração dos dados que foram estruturados em esquemas distintos. Ele comenta que um ponto a ser considerado é o espaço de solução que é considerado vasto.

Todo este processo auxilia na centralização dos dados, e que, como é elaborado por Correa (2008).

> A centralização desses dados é de suma importância, pois reduz esforços na obtenção de dados de grandes repositórios, permitindo que esses esforços sejam dispendidos na análise na tomada de decisão, ou seja, retirar informações dos dados. (CORREA, 2008)

Tarefa essa que acaba sendo resolver a questão inicial dos usuários, pois afinal, o desejado é a informação final, não o seu processo de filtragem.

1.3 Metodologia

A metodologia desta pesquisa contará com várias etapas, como demonstra o Tabela 1 que apresentam as etapas do estudo ao longo dos semestres. Esta abordagem visa garantir uma progressão consistente e focalizado nos desafios inerentes a cada fase do projeto nesta área de pesquisa. Primeiro, profissionais e acadêmicos da área serão entrevistados para se entender em termos práticos os desafios recorrentemente encontrado, em seguinda será realizada uma revisão bibliográfica para esclarecer aprofundadamente os desafios e soluções apresentadas pelos especialistas.

1.3.1 Desenvolvimento do Sistema de Web Scraping

O desenvolvimento do sistema de web scraping será dividido em várias fases:

Coleta de Dados: Implementação de scripts para a coleta de dados de diferentes fontes web, utilizando bibliotecas como BeautifulSoup, Scrapy ou Selenium. Tratamento e Limpeza dos Dados: Desenvolvimento de algoritmos para a limpeza e normalização dos dados coletados, garantindo a consistência e a qualidade das informações. Armazenamento dos Dados: Utilização de estruturas de dados adequadas (como bancos de dados relacionais e NoSQL) para o armazenamento eficiente dos dados coletados.

1.3.2 Aplicação de Técnicas de Mineração de Dados

Nesta etapa, serão aplicadas técnicas de mineração de dados para extrair informações relevantes dos dados coletados:

Análise de Texto: Utilização de algoritmos de NLP para extrair informações textuais relevantes. Clusterização e Classificação: Aplicação de métodos de aprendizado de máquina para agrupar e classificar os dados, facilitando a identificação de padrões e tendências.

1.3.3 Desenvolvimento do Sistema de Visualização de Dados

Com base nas informações extraídas, será desenvolvido um sistema de visualização de dados que permita ao usuário final interagir de maneira eficiente com os dados:

Desenvolvimento da Interface: Criação de uma interface web intuitiva, utilizando frameworks como React ou Angular. Visualizações Interativas: Implementação de gráficos, tabelas e outras visualizações interativas que permitam ao usuário explorar os dados de maneira dinâmica. Filtros e Ordenações: Desenvolvimento de funcionalidades que permitam ao usuário filtrar e ordenar os dados conforme suas necessidades.

1.3.4 Integração e Testes

Após o desenvolvimento dos módulos individuais, será realizada a integração do sistema completo. Esta etapa incluirá:

Testes Unitários e de Integração: Aplicação de testes para garantir que todos os componentes do sistema funcionem corretamente e de forma integrada. Testes de Usabilidade: Realização de testes com usuários reais para avaliar a usabilidade e a eficiência da interface de visualização.

1.3.5 Validação e Avaliação do Sistema

Por fim, o sistema será validado e avaliado com base nos critérios definidos na etapa de requisitos:

Desempenho: Medição do tempo de resposta e eficiência do sistema. Qualidade dos Dados: Avaliação da precisão e da relevância das informações extraídas. Satisfação do Usuário: Coleta de feedback dos usuários para identificar possíveis melhorias e ajustar o sistema conforme necessário.

1.4 Cronograma

O cronograma de atividades é apresentado na Tabela 1 demarca o período temporal em que estima-se que as atividades serão realizadas. Elas estão divididas em três categorias por similaridade da atividade: disciplinas **(DC)**, estudo da literatura **(LT)** e atividades gerais (**GR**). As disciplinas são divididas em quatro categorias, cada uma com um número sugerido de disciplinas a serem cursadas, sendo elas: núcleo comum (2), interesse do aluno (3), demais áreas (5) e tópicos especiais (sem quantidade sugerida). No cronograma, as disciplinas estão agrupadas por blocos de similaridade, sendo seguidas pelo respectivo estudo da literatura e por fim as atividades gerais.

Tabela 1 – Cronograma de atividades

Atividade	Semestres 1	2	3	4
DC: Teoria da Informação				
DC: Recuperação de Informação				
LT: Visualização de informação				
DC: Mineração de Dados				
DC: Visualização de Dados				
DC: Processamento de Dados Massivos em Nuvem				
DC: Aprend. Profundo p/ Proc. de Linguagem Natural				
LT: Mineração de dados				
DC: Inteligência Artificial				
DC: Tóp. em Inteligência Artificial				
LT: Inteligência Artificial				
DC: Bancos de Dados				
DC: Tóp. em Bancos de Dados				
DC: Projeto e Análise de Algoritmos*				
DC: Programação Paralela				
DC: Fund. Teór. da Comp.				
LT: Sistemas similares				
DC: Teoria dos Grafos				
LT: Organização do conhecimento				
DC: Tarefas ou estudos especiais				
GR: Desenvolvimento do sistema				
GR: Escrita da dissertação				

REFERÊNCIAS

ARANTES, A. R. Web View: Uma Ferramenta para Construção de Visões de Fontes de Dados da Web. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, jul. 2001. Disponível em: http://hdl.handle.net/1843/BUBD-9BQK6L. Acesso em: 30/05/2024. Citado na página 1.

CARVALHO, M. G. D. EVOLUTIONARY APPROACHES TO DATA INTEGRATION RELATED PROBLEMS. Tese (Doutorado) — Universidade Federal de Minas Gerais, out. 2009. Citado na página 2.

CORREA, M. D. Web2DB - Uma Ferramenta para Construção de Representações Relacionais de Sítios da Web. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, mar. 2008. Citado 2 vezes nas páginas 1 e 2.

FORTES, R. S. Enhancing the Multi-Objective Recommendation from three new perspectives: data characterization, risk-sensitiveness, and prioritization of the objectives. Tese (Doutorado) — Universidade Federal de Minas Gerais, maio 2022. Citado na página 2.

SILVA, A. S. da. Estratégias Baseadas em Exemplos para Extracão de Dados Semi-Estruturados da Web. Tese (Doutorado) — Universidade Federal de Minas Gerais, jun. 2002. Disponível em: http://hdl.handle.net/1843/SLBS-5KKKXX. Acesso em: 30/05/2024. Citado na página 1.

VENEROSO, J. M. D. F. Reconhecimento de Entidades Nomeadas na Web. Dissertação (Mestrado) — Universidade Federal de Minas Gerais, ago. 2019. Citado na página 2.