WIKIPÉDIA GEOHIST: UMA FERRAMENTA DE PESQUISA BASEADA NA WIKIPÉDIA PARA GERAÇÃO DE VISUALIZAÇÕES GRÁFICAS INTERATIVAS

Fabrício da Silva Rocha*
Ricardo Lenz César**
Diego Rocha Lima***

RESUMO

Na sociedade atual, a crescente demanda por atividades envolvendo busca, coleta e análise de dados na aquisição de conhecimentos traz consigo desafios de como se orientar em meio a uma abundância de conteúdos. Nesse contexto, diversas ferramentas são utilizadas para tornar o processo de aprendizagem e aquisição de conhecimentos mais eficiente e com menor sobrecarga cognitiva para os usuários. Visualizações gráficas interativas, em especial, facilitam bastante esses processos. Considerando que a Wikipédia, hoje, é um dos meios mais utilizados como base de informações na internet, com abundância de artefatos e a diversidade de mídias e idiomas, o presente trabalho se enquadra nas ferramentas de soluções digitais que facilitam usufruir do potencial dessa base pública e universal de informações, mediante utilização de visualizações gráficas interativas voltadas para alunos, professores e demais interessados na construção do seu conhecimento. O trabalho propõe uma ferramenta que busque e apresente informações de maneira gráfica, dispondo os resultados em um mapa com linha de tempo cronológica, fomentando um contexto geográfico e histórico de maneira visual para exploração dos conteúdos pesquisados por meio dela, assim como do histórico de navegação do usuário, ajudando-o com estatísticas sobre suas próprias buscas. As informações são dispostas numa composição em comum, permitindo a interconexão de conteúdos diversos numa estrutura uniforme, e sem a necessidade do usuário ter que lidar com os formatos particulares de como a Wikipédia organiza suas informações de cunho geográfico e histórico. A ferramenta está disponível como software livre, para que mais pessoas da comunidade possam livremente utilizar. Além disso, a possibilidade de uso por desenvolvedores também é considerada por meio de um *framework* interno, disponibilizado para atividades em torno do tema.

Palavras-chave: Conhecimento. Visualizações Gráficas. Wikipédia.

* Graduando em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Ceará, Brasil. E-mail: fabriciosilvalp@outlook.com
** Mestre em Ciência da Computação pela Universidade Federal do Ceará — UFC, Docente do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Ceará, Brasil. E-mail: ricardo.lenz@ifce.edu.br
*** Doutor em Engenharia da Computação pela Universidade Federal do Rio Grande do Norte — UFRN, Docente do Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE), Aracati, Ceará, Brasil. E-mail: diego.rocha@ifce.edu.br

ABSTRACT

In today's society, the growing demand for activities involving data search, collection, and analysis in knowledge acquisition brings challenges on how to navigate through an abundance of content. In this context, various tools are used to make the learning and knowledge acquisition process more efficient and less cognitively burdensome for users. Interactive graphical visualizations, in particular, greatly facilitate these processes. Considering that Wikipedia is now one of the most widely used sources of information on the internet, with an abundance of artifacts and diverse media and languages, this work falls into the category of digital solutions tools that facilitate harnessing the potential of this public and universal knowledge base through the use of interactive graphical visualizations aimed at students, teachers, and other individuals interested in constructing their knowledge.The work proposes a tool that seeks and presents information in a graphical manner, displaying the results on a map with a chronological timeline, fostering a visual geographic and historical context for exploring the researched content, as well as the user's browsing history, providing statistics on their own searches. The information is presented in a unified composition, allowing for the interconnection of various contents in a consistent structure, without the user having to deal with the specific formats in which Wikipedia organizes its geographic and historical information. The tool is available as Open-Source software, so that more people in the community can freely use it. Furthermore, the possibility of use by developers is also considered through an framework provided for activities around the theme.

Palavras-chave: Knowledge. Graphical Visualization. Wikipedia.

1 INTRODUÇÃO

Na sociedade atual, na assim chamada era da informação (JAMIL; NEVES, 2000), o trabalho envolvendo busca, coleta e análise de dados é cada vez mais demandado por estudantes de várias idades e profissionais dos mais diversos ramos de atuação. Nessa era da informação, onde há amplo uso de tecnologias da informação e comunicação (TICs), várias pessoas trabalham com serviços onde ocorre intensa atividade ligada à informação, sendo a aquisição e o trato das mesmas um fator crucial para seus dinamismos, trazendo impacto em várias áreas, desde negócios até a educação. A informação pode fazer grande diferença para muitas pessoas se elas puderem acessá-la e inspecioná-la efetivamente.

Um desafio enfrentado na atualidade está em como propiciar conhecimento a partir de um volume tão grande de informações. Muitas pessoas, por meio de um celular conectado à internet, podem conseguir rapidamente acessar plataformas volumosas de dados. Em diversos lugares, a disponibilidade púbica de informações é muito mais ampla hoje do que no passado; por outro lado, a abundância de conteúdo oferecido levanta um desafio à parte, a saber, como se orientar num mar de informações e compreender de forma mais eficiente, menos custosa

aquilo que esta buscando. Com efeito, a informação pode estar disponível publicamente, mas não se torna conhecimento enquanto não é compreendida, interconectada, apreendida — daí a importância de mecanismos que viabilizem esse caminho, melhorando a capacidade das pessoas de digerirem o que recebem.

A fim de pavimentar um caminho mais facilitado para esse tipo de tarefa, ferramentas digitais diversas têm sido utilizadas visando tornar o processo de aprendizagem e aquisição de conhecimento mais eficiente e com menor sobrecarga de esforço cognitivo. Particularmente, visualizações gráficas podem ser de grande ajuda no processo de aquisição de conhecimento sem demandar um empreendimento mental mais intenso por parte dos usuários, podendo enriquecer bastante esse processo.

Na atualidade, o Buscador do Google (JAMALI; ASADI, 2010) e a enciclopédia da Wikipédia (WALLACE; FLEET, 2005) têm sido algumas das ferramentas mais amplamente utilizadas no mundo para a busca de informações. A Wikipédia é hoje a maior enciclopédia do mundo, com grande quantidade de dados e presente em inúmeros idiomas. Algumas soluções desenvolvidas a partir dessas ferramentas têm buscado extrair mais possibilidades para os usuários. Em particular, a ideia de buscar dados da Wikipédia e apresentá-los numa disposição gráfica envolvendo mapa e sequência temporal permite atender diversos segmentos, desde alunos, professores e demais interessados numa exploração histórica e geográfica de pessoas, movimentos, instituições e outros tópicos da enciclopédia.

A visualização de informações por meio de gráficos interativos permite certos níveis de exploração por parte do usuário que podem ser estímulos importantes para várias pessoas. Uma apresentação contendo resultados estatísticos permite, inclusive, buscar contrapontos e complementos em relação ao que foi pesquisado. Por exemplo, um usuário pode identificar de forma mais fácil que, ao longo de vários anos de buscas sobre tópicos enciclopédicos, acabou concentrando-se demais em um período ou região específica.

Assim, o presente trabalho propõe uma ferramenta que busque informações desejadas pelo usuário e as apresente num gráfico interativo, utilizando como fonte de pesquisa a Wikipédia. O trabalho propõe isso na linha anteriormente mencionada, de buscar e apresentar informações numa disposição gráfica que demande menor esforço cognitivo de compreensão, facilitando assim o processo de aprendizagem. Ademais, será possível a composição de diversos elementos de pesquisa num só resultado, apresentando os diversos tópicos sendo buscados num mesmo mapa espacial-temporal, permitindo a conexão de ideias de maneira criativa por parte dos usuários com o intuito de incentivar a livre exploração de tópicos na pesquisa.

A ferramenta permitirá a extração desse tipo de informação sem exigir do usuário um conhecimento mais detalhado das diferentes formas de organização de dados da Wikipédia, diminuindo a lacuna entre o que o usuário poderia mais facilmente obter e o que hoje muitos ainda não têm. Além disso, também estão disponíveis funcionalidades com a capacidade de leitura e extração de informações a partir do histórico de navegação do usuário, permitindo também a geração de visualizações, de tal maneira que estas possam ser usadas para, dentre diversas outras finalidades, traçar um perfil do usuário e revelar tendências de buscas na Wikipédia,

por exemplo. Por fim, será possível gerar um documento compilado com todas as informações inseridas pelo usuário, bem como o produto gerado por essas entradas, como a lista dos tópicos pesquisados, linha histórica e mapa de regiões, permitindo o compartilhamento com outras pessoas, e facilitando a troca de informações, podendo servir como fonte bibliográfica para usos futuros.

A utilidade da ferramenta compreende não apenas a extração de dados da Wikipédia, mas reside no produto resultante desse processo que encapsula informações relevantes e apresenta visualizações significativas, permitindo ao usuário compreender a distribuição de suas pesquisas no mapa, no tempo. A ferramenta pode fornecer *insights* valiosos sobre os usuários e seu horizonte literário. Embora a Wikipédia seja a fonte inicial dos dados, é a transformação e a interpretação desses dados que agregam valor.

O trabalho está disponibilizado como um software livre, permitindo assim seu uso prático e imediato por parte da comunidade. Espera-se contribuir, assim, para alunos, professores e demais interessados no tema, facilitando o processo de compreensão de certas informações em suas pesquisas e agregando maior produtividade para os usuários.

O presente trabalho está organizado da seguinte forma: na Seção 2 apresenta-se a fundamentação teórica, fazendo-se uma exposição da Wikipédia e como seus dados são organizados, além de descrever como lidar com os dados do histórico de navegação do usuário; na Seção 3, são apresentados os trabalhos relacionados, tanto artigos quanto implementações relevantes, com tabelas comparativas e discussão sobre eles; na Seção 4 a proposta do trabalho é apresentada, com uma visão geral do sistema implementado; na Seção 5 são mostrados os resultados obtidos, as telas do sistema, os gráficos gerados, etc; por fim, a Seção 6 traz as conclusões e sugestões de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 A WIKIPÉDIA E SEUS DADOS

Surgida em 2001, na sucessão de outro projeto dos mesmos criadores — *Nupedia* (SANGER, 2005) — a Wikipédia é uma enciclopédia multilíngue, online, de licença livre e de escrita colaborativa, ou seja, os artigos são criados e editados pela própria comunidade, que contribuem com seu tempo e conhecimento para manter a qualidade e a precisão das informações disponíveis na enciclopédia. Criada por Jimmy Wales e Larry Sanger, a Wikipédia é distribuída em cerca de 300 idiomas diferentes[1], alguns mais ativos que outros (303 ao total, mas 291 ativos), e contém hoje, milhões de artigos. A versão em inglês, por exemplo, conta com aproximadamente 6.6 milhões de artigos[2], e em português, pouco mais de 1 milhão[3]. Os artigos podem ser editados, adicionando ou revisando o conteúdo existente, citando fontes e referências, enriquecendo cada vez mais a informação. O nome Wikipédia deriva da fusão de

[1] https://pt.wikipedia.org/wiki/Lista_de_Wikipédias
[2] https://en.wikipedia.org/wiki/Wikipedia:Size_comparisons
[3] https://pt.wikipedia.org/wiki/Wikipédia:Estatísticas

duas palavras: *wiki*, um termo havaiano que significa rápido ou ágil, e *encyclopedia*, que significa enciclopédia em inglês. Jimmy Wales escolheu o nome porque queria transmitir a ideia de uma enciclopédia online que fosse rápida e fácil de usar. A partir de 2003, a Wikipédia passou a ser mantida e gerida pela *Wikimedia Foundation*[4], uma organização sem fins lucrativos fundada pelos mesmos criadores da Wikipédia. A fundação objetiva incentivar a produção, desenvolvimento e distribuição de conteúdo livre e em diversos idiomas, e também é responsável por manter outros projetos semelhantes à Wikipédia, como o Wikcionário, Wikilivros e *Wikidata*. Desde sua fundação, ela desempenha um papel importante na promoção do acesso gratuito à informação e no fortalecimento da cultura colaborativa e de compartilhamento de conhecimento na internet.

A Wikipédia, que se tornou referência na busca de informações e conhecimento na internet (KERN, 2018), assim como outros projetos da fundação a qual pertence, é baseada na tecnologia *Wiki* (RAMAN, 2010), um software livre e gratuito para gerenciamento de conteúdo e colaboração online, criado em 1995 pelo programador americano Ward Cunningham. O software *Wiki* permite que qualquer pessoa possa criar e editar conteúdo, sem a necessidade de habilidades técnicas avançadas.

O *Wiki* é uma espécie de sistema de gestão de conteúdo, que permite a criação de páginas *web* colaborativas de maneira simples e intuitiva. Através dele, é possível criar páginas, adicionar texto, imagens, *links*, e editar o conteúdo existente de maneira conjunta. Dessa forma, a Wikipédia utiliza-o para permitir que pessoas do mundo inteiro possam contribuir coletivamente na criação de uma enciclopédia global e gratuita.

Isso significa que qualquer pessoa com acesso à internet pode contribuir para a criação e manutenção de conteúdo na Wikipédia, por exemplo. Utiliza a linguagem wikitexto, uma linguagem de marcação intermediária que permite aos usuários formatar o conteúdo das páginas em sistemas *wiki* de forma simples, adicionando cabeçalhos, listas, *links* e estilos de texto. Essa linguagem facilita a formatação rápida e intuitiva do texto, sem a necessidade de conhecimentos avançados de codificação, sendo convertida em *HTML* pelo software *wiki* para visualização nos navegadores *web*. A sintaxe do wikitexto pode variar entre diferentes sistemas *wiki*.

Além disso, também é utilizado para criar outras *wikis*, tanto pela *Wikimedia Foundation*, como por outras organizações e indivíduos, para os mais diversos fins, como criação de guias, documentação técnica, fóruns de discussão, entre outros. A Wikipédia utiliza uma implementação mais avançada e sofisticada do conceito de wikis originalmente desenvolvido, chamada *MediaWiki*.

Com o grande volume de informações disponíveis, a Wikipédia constitui uma rica fonte de conhecimento e um recurso com potencial para diversas áreas de pesquisa, como processamento de linguagem natural, gerenciamento de conhecimento, mineração de dados e outras (MILNE; WITTEN, 2013). Uma forma genérica de se obter dados da Wikipédia seria via *Web Scraping*, uma técnica de coleta de dados de uma página *web* de forma automatizada. Nesse cenário, é possível desenvolver uma ferramenta com a função de extrair dados de uma página com

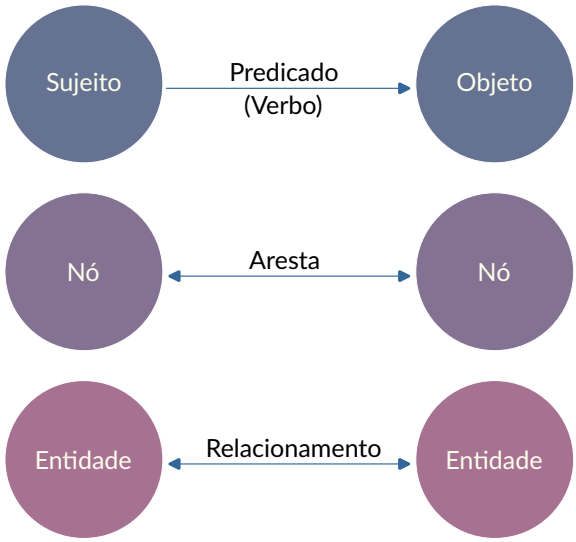
[4] https://wikimediafoundation.org/

rapidez e precisão. As ferramentas podem variar na complexidade, a depender de onde o usuário deseja extrair as informações. Parte importante da ferramenta está nos localizadores ou seletores de dados usados para encontrar as informações desejadas na página, estruturada em *HTML*. Podem ser utilizados *XPath* (OLTEANU et al., 2002), seletores CSS, *tags HTML*, expressões regulares, ou até mesmo uma combinação deles. Diversas linguagens de programação suportam o desenvolvimento de ferramentas para *Web Scraping*.

Alternativamente, essa extração pode ser feita através da DBpédia (AUER et al., 2007), um empenho da comunidade de *crowdsourcing* para extrair conteúdo de forma estruturada das informações criadas em várias fontes de dados, porém com mais significância da Wikipédia. Ela extrai informações estruturadas de artigos em diferentes idiomas e as converte em um conjunto de dados RDF (*Resource Description Framework*), uma forma de representar dados na *web* Semântica por meio de grafos. Com isso, os dados da DBpédia podem ser usados por outras aplicações e sistemas para gerar conhecimento e *insights* a partir do vasto conjunto de informações disponíveis na Wikipédia. Outra forma seria através da *Wikidata*, um projeto da *Wikimedia Foundation* lançado em 2012 que visa a criação de uma base de dados estruturada e colaborativa para o armazenamento de informações. Apesar das semelhanças, DBpédia e *Wikidata* possuem diferenças significativas. Em resumo, a DBpédia é mais focada em extrair informações estruturadas da Wikipédia, já a *Wikidata* visa criar uma base de conhecimento livre e aberta, que possa ser utilizada por pessoas e máquinas para diversos fins. Ela atua como ponto central de armazenamento para os dados estruturados de projetos sobre o domínio da *Wikimedia Foundation*, incluindo Wikipédia, Wikiversidade, Wikcionário, Wikilivros, entre outros.

Um ponto em comum entre ambas as iniciativas é a utilização do modelo RDF para representação de dados. O modelo RDF (MCBRIDE, 2004) é composto por três componentes básicos: sujeito, predicado e objeto, os quais também podem ser encontrados em algumas literaturas como recurso, propriedade e valor, respectivamente, ou simplesmente *tripla*. Esses componentes são usados para descrever recursos e suas relações na forma de grafos, onde os recursos são representados por nós e as relações por arestas. O sujeito é um recurso que está sendo descrito e é representado por um nó no grafo. Já predicado é a propriedade que está sendo atribuída ao sujeito, representado pela aresta que liga o nó do sujeito ao nó do objeto. Por fim, objeto é o valor da propriedade, que pode ser outro recurso ou um literal, como uma cadeia de caracteres ou um número, também visualizada como um nó no grafo. A figura abaixo ilustra os componentes e suas relações.

Figura 1 – Estrutura e relações nas triplas RDF



Fonte: Elaborado pelos autores.

Essa maneira de estruturar os dados proporciona diversos benefícios, entre eles escalabilidade ao sistema. Por exemplo, se quisermos descrever que “João é filho de Maria”, podemos representar isso no modelo RDF da seguinte forma:

- **Sujeito**: João;
- **Predicado**: é filho de;
- **Objeto**: Maria.

Mediante uma linguagem específica chamada *SPARQL* (PÉREZ; ARENAS; GUTIERREZ, 2006), é possível fazer consultas complexas em bancos de dados RDF. O termo *SPARQL* é um acrônimo recursivo que significa *SPARQL Protocol and RDF Query Language*. É similar em sintaxe e funcionalidade à linguagem *SQL* usada em bancos de dados relacionais, mas o foco é em dados na *web* e em bancos de dados semânticos. O *SPARQL* permite que os usuários façam consultas complexas utilizando padrões de triplas, operadores lógicos e outras construções para especificar que informações desejam obter. É possível realizar consultas que envolvem vários grafos RDF, unir informações de diferentes fontes e obter resultados em diversos formatos, como *JSON*, *CSV*, *XML*, entre outros. Trata-se de uma ferramenta poderosa para lidar com dados na *web*, possibilitando a recuperação de informações de maneira mais eficiente e precisa do que outras formas de busca. Tanto a DBpédia quanto a *Wikidata* disponibilizam formas de realizar consultas *SPARQL* online.

2.2 APLICAÇÕES WEB COM GRÁFICOS INTERATIVOS

Aplicações *web* são programas ou softwares com implementações que recorrem ao uso de tecnologias com *HTML*, *CSS* e *Javascript*, entre outras. Elas podem ser executadas via navegador, sem a necessidade de instalação local, assim como a partir de servidores *web* configurados no equipamento do próprio usuário ou em servidores de terceiros. Geralmente, essas aplicações seguem o modelo cliente/servidor, em que o servidor aguarda uma requisição do usuário via protocolo *HTTP*. Após o processamento da requisição, o servidor envia um pacote de dados em resposta ao cliente, que pode conter uma página *HTML*, objetos *JSON*, *XML*, ou outros formatos.

Nesse contexto, muitas vezes é necessário gerar conteúdos e arquivos de forma dinâmica para apresentar informações de maneira mais clara e interativa para o usuário. Para atender a essa necessidade, existem diversas bibliotecas e *frameworks* visuais disponíveis que podem ser incorporados ao desenvolvimento de aplicações *web*. Entre essas bibliotecas, é possível destacar as que oferecem suporte para visualizações geoespaciais, como a Folium[5] para *Python*, que utiliza o poder de manipulação de dados dessa linguagem para gerar mapas interativos com o auxílio da biblioteca para a linguagem *Javascript* chamada *Leaflet*[6]. Além disso, as bibliotecas *Chart.js*[7] e *D3*[8] para *Javascript*, e ainda *Matplotlib*[9] e *Seaborn*[10] para *Python*, podem ser usadas para gerar os mais variados gráficos, inclusive gráficos de bolhas, como, por exemplo, o que é apresentado mais adiante no texto, na Figura 8 (descrita também mais à frente). Essas bibliotecas podem ajudar a criar visualizações mais atrativas e dinâmicas, além de possibilitar uma melhor compreensão dos dados pelos usuários finais.

2.3 BANCOS DE DADOS DO HISTÓRICO DE NAVEGAÇÃO

Os navegadores, como *Google Chrome*, *Firefox* e outros, costumam armazenar o histórico de pesquisas do usuário como um banco de dados em arquivo *Sqlite* (BHOSALE; PATIL; PATIL, 2015). Este histórico pode ser utilizado como fonte de dados para geração de visualizações e *insights*, além de permitir traçar um perfil do usuário em particular. As informações armazenadas incluem *URLs* visitadas, informações sobre downloads de arquivos e palavras-chave pesquisadas. Trabalhar com o histórico do usuário implica em trabalhar com consultas a esse banco de dados. Uma das possibilidades é saber quais as variantes da Wikipédia mais acessadas (por exemplo, inglês, português) e a quantidade de acessos de cada uma. Além disso, é possível revelar tendências de busca a temas específicos, horários do dia em que as buscas foram realizadas, sites acessados antes do acesso atual, e assim apresentar o trajeto feito até chegar à página em questão.

Por exemplo, com o navegador Mozilla Firefox, no Linux, tipicamente os dados de perfil do usuário ficam em */.mozilla/firefox/*. Nesse diretório, um arquivo chamado *profiles.ini* descreve

[5] https://python-visualization.github.io/folium/
[6] https://leafletjs.com/
[7] https://www.chartjs.org/
[8] https://d3js.org/
[9] https://matplotlib.org/
[10] https://seaborn.pydata.org/

múltiplos perfis, inclusive especificando qual é o padrão (*default*) e em que subdiretório ele se encontra. Uma vez no subdiretório correto, o arquivo *places.sqlite* contém o banco de dados em questão, com o qual pode-se, por exemplo, executar consultas *SQL* como a apresentada no código abaixo. Neste exemplo, a consulta mostra que, no *Firefox*, uma tabela contém os lugares acessados e outra contém o histórico das visitas a tais lugares; nesse histórico, pode-se obter a data de acesso e até, inclusive, a hora.

```
SELECT
    h.visit_date, p.url

FROM
    moz_historyvisits h JOIN moz_places p ON h.place_id = p.id

WHERE
    INSTR(p.url, "/en.wikipedia.org") > 0 OR
    INSTR(p.url, "/pt.wikipedia.org") > 0

ORDER BY
    h.visit_date DESC;
```

Exemplo de consulta ao histórico do usuário

3 TRABALHOS RELACIONADOS

Nesta seção, são apresentados trabalhos inseridos no contexto da exploração da Wikipédia através da extração de informações de maneira automatizada, assim como a geração de visualizações gráficas desses dados. Dentre os trabalhos apresentados estão publicações científicas, ferramentas online e repositórios no *Github*.

Em Utomo (2013) o autor propõe uma ferramenta que utiliza a técnica de *Web Scraping* e o uso de expressões regulares para extrair automaticamente o conteúdo principal de uma página da *web*, nesse estudo, a Wikipédia. O sistema tem como base um programa, escrito na linguagem *PHP*, responsável pela busca das páginas, por meio de palavras-chave, e pela extração dos seus dados. A ferramenta é incorporada na forma de um *plugin* à plataforma de gerenciamento do conteúdo *Wordpress*, a qual é utilizada para gerenciar e exibir o conteúdo extraído na forma de artigos na plataforma. Funções de busca por artigos anteriormente extraídos e armazenados na plataforma também são implementadas. Nenhuma análise mais aprofundada ou geração de visualização dos dados é implementada, apenas a exibição dos dados extraídos.

Já em Biuk-Aghai, Pang e Si (2014), a fim de propor um estudo sobre a colaboração humana voluntária em grande escala na *web* 2.0, mas especificamente no contexto da Wikipédia, os autores analisam detalhadamente a coautoria de artigos em toda Wikipédia, em diversos idiomas, revelando padrões que seguem uma distribuição geométrica. Para uma melhor compreensão dos padrões encontrados, os dados analisados foram categorizados e dispostos de forma

visual, semelhante a um mapa geográfico. A visualização gerada apontou diversas diferenças significativas na contagem de coautores em diferentes tópicos em todas as edições de idioma da Wikipédia analisadas, diferenças detalhadas no decorrer do trabalho.

Devido à natureza semiestruturada do conjunto de dados que compõe a Wikipédia, e também sua grande dimensão, a integração dos dados em tabelas estruturadas para posterior visualização se torna um desafio à parte. Chan et al. (2008) propõe a *Vispedia*, um sistema de visualização baseado na *web* que busca minimizar esse problema. A ferramenta permite a livre navegação na Wikipédia, possibilitando que o usuário possa escolher tabelas de seu interesse e, por meio de interface visual, adicione colunas a sua tabela inicial, compondo assim, novas visualizações. A aquisição das informações é feita através de algoritmos executados sobre a DBpédia.

Com o foco em eventos históricos, como guerras, batalhas e invasões, Chasin (2010) explora a tarefa de criar uma linha temporal desses eventos. Seu esforço é concentrado em obter apenas os eventos mais importantes associados a uma data, tendo em vista que as ferramentas existentes apenas listam os eventos de maneira geral, sem essa classificação. Após a coleta e classificação das informações, uma visualização na forma de linha do tempo é gerada, via uma interface *web*, onde são exibidas informações como descrições e nomes daqueles eventos obtidos, suas datas e localizações no mapa.

Conforme citado anteriormente, a DBpédia extrai informações de forma estruturada da Wikipédia e as disponibiliza na *web*, tornando as informações legíveis por humanos e máquinas. As informações usadas para popular a DBpédia são retiradas, principalmente das *infoboxes* dos artigos na Wikipédia, caixas com informações resumidas, quando disponíveis, usadas para concentrar e estruturar algumas informações do artigo. Porém, muitas informações que se encontram fora dessas *infoboxes*, ou seja, no corpo do artigo, não são mapeadas para a estrutura da DBpédia. Tendo isso em mente, Hienert e Luciano (2012) propõe a extração de informações sobre eventos históricos de artigos da Wikipédia, concentrando-se nos casos que não são mapeados para a DBpédia, no intervalo de 2.500 anos, em diferentes idiomas. Os dados obtidos são disponibilizados por meio de uma *API web*, e modelados no formato RDF, tornando-os legíveis à DBpédia, e permitindo serem consultados através da linguagem *SPARQL*.

Russo, Caselli e Monachini (2015) propõem o desenvolvimento de um módulo de PLN (Processamento de Linguagem Natural) para análise, extração e visualização de eventos temporais em biografias disponíveis na Wikipédia. A obtenção das informações é baseada em datas, fato esse que pode ser usado para identificar outros eventos biográficos relevantes, facilitando a ordenação cronológica e criando uma cadeia de eventos, fornecendo uma linha histórica pessoal.

No trabalho de Sipoš et al. (2009) é descrita a *HistoryViz*, ferramenta que permite o usuário, ao invés de navegar pelos artigos da Wikipédia da maneira tradicional, visualizar as relações que conectam diferentes eventos e entidades selecionadas. Por meio de uma interface gráfica, é possível visualizar na forma de uma linha do tempo os eventos referentes a entidade escolhida, bem como gráficos que mostram a relação dela com outras entidades, com a capacidade de se expandirem dinamicamente.

Merece destaque ainda o trabalho de Wang et al. (2010), que estendendo algumas funcionalidades presentes na base de conhecimento *YAGO* (REBELE et al., 2016), com aspectos e características temporais, apresenta a ferramenta *Timely YAGO*. Ainda em protótipo, ela propõe a extração de fatos temporais das *infoboxes*, categorias e listas presentes no artigo e integra na sua própria base de conhecimento. Além disso, a ferramenta fornece suporte para consultas, com viés temporal, na linguagem *SPARQL*, uma vez que os dados são estruturados no modelo RDF. Abaixo, uma tabela comparativa [1] entre a proposta apresentada e os trabalhos aqui relacionados.

Tabela 1 – Tabela comparativa

Trabalho	**Web Scraping**	**Wikidata**	**Histórico de navegação**	**Visual.**
(HIENERT; LUCIANO, 2012)	-	-	-	OK
(UTOMO, 2013)	OK	-	-	-
(BIUK-AGHAI; PANG; SI, 2014)	-	-	-	OK
(CHAN et al., 2008)	-	-	-	OK
(CHASIN, 2010)	-	-	-	OK
(RUSSO; CASELLI; MONACHINI, 2015)	-	-	-	OK
(SIPOŠ et al., 2009)	-	-	-	OK
(WANG et al., 2010)	-	-	-	OK
Wikipédia GeoHist[11]	OK	OK	OK	OK

Vários trabalhos lidam com os dados por meio da plataforma *Wikidata* e estão listados oficialmente no próprio site da plataforma[12], subdivididos em categorias diversas como melhorias de interface do usuário (*enhance user interface*), consulta de dados (*query data*) e visualização de dados (*visualize data*), com uma categoria ainda mais ligada para desenvolvedores (*for programmers*). Na categoria de ferramentas focadas em visualização de dados, pode-se citar as seguintes ferramentas:

Tabela 2 – Ferramentas online e repositórios

	Projetos	**Descrição / Funcionalidade**
(1)	Histomania[13]	Aplicação *web* para publicação de histórias, oferecido em alemão, com visualização de eventos em linhas de tempo e mapas geográficos a partir da *Wikidata* e Wikipédia, mirando um público que tenha interesse em história.

[11] https://github.com/faahiero/TimelineAppV2
[12] https://www.wikidata.org/wiki/Wikidata:Tools
[13] https://histomania.com/

	Projetos	Descrição / Funcionalidade
(2)	Histropedia[14]	Usa dados da Wikipédia e *Wikidata* para gerar linhas de tempo interativas com eventos vinculados a artigos da Wikipédia, permitindo também ao usuário criar sua própria linha de tempo.
(3)	EntiTree[15]	Visualização gráfica de uma árvore genealógica de pessoas à partir da *Wikidata*.
(4)	Denelezh[16]	Estatísticas e visualização gráfica sobre diferenças de gênero nos registros.
(5)	GeneaWiki[17]	Para visualização de genealogias, semelhante ao EntiTree.
(6)	Linked People[18]	Semelhante ao EntiTree, apresenta genealogias entre personagens de filmes, livros, etc. à partir da *Wikidata*.
(7)	Missing images heatmap[19]	Visualização mediante um mapa de calor para itens da *Wikidata* que não estão com imagens ainda.
(8)	Wikipedia-Scraper[20]	Dado um verbete, extrai a imagem associada
(9)	Wiki-table-scrape[21]	Dado um *link* e um cabeçalho, extrai informações de uma tabela para o formato *CSV*.
(10)	Wiki-parser-py[22]	Dado um *link*, realiza scraping de informações diversas da Wikipédia, incluindo o infobox em formato *JSON*.
(11)	Wikipedia-crawler[23]	Encontra caminho entre duas páginas da Wikipédia.
(12)	Wiki-graph[24]	Ferramenta de comando e aplicação *web* para gerar um grafo visual de conexões a partir de uma busca.
(13)	Archaea[25]	Ferramenta que gera um grafo visual de conexões a partir de uma busca.
(14)	Wikipedia-map[26]	Aplicação *web* para visualizar conexões entre páginas da Wikipédia de uma forma gráfica.

[14] http://histropedia.com/timeline/
[15] https://www.entitree.com/
[16] https://denelezh.wmcloud.org/
[17] https://magnus-toolserver.toolforge.org/ts2/geneawiki/
[18] https://linkedpeople.net/
[19] https://wikidata-todo.toolforge.org/missing_images.html
[20] https://github.com/hmnhGeek/Wikipedia-Scraper
[21] https://github.com/rocheio/wiki-table-scrape
[22] https://github.com/KiranNiranjan/wiki-parser-py
[23] https://github.com/theimberger/wikipedia_crawler
[24] https://github.com/francis-du/wiki-graph
[25] https://github.com/kayandrewj/Archaea
[26] https://github.com/controversial/wikipedia-map

Várias das ferramentas oferecidas são mais voltadas para facilidades de extração de dados da Wikipédia, como os trabalhos (8), (9), (10) e (11). Essas implementações são indicadas a desenvolvedores, e podem ser modificadas para atender demandas específicas. Por outro lado, outros projetos são trabalhos relacionados mais próximos que oferecem uma interface de usuário para explorar o conteúdo, como os trabalhos (1) e (2), permitindo a criação de visualizações como linhas do tempo. Algumas abordagens focam em estatísticas sobre os dados da Wikipédia, como o trabalho (4). Os trabalhos (3), (5), (12), (13) e (14) permitem explorar conexões de um verbete como que num grafo visual, contudo, exploram apenas as conexões em si, sem estabelecer uma exploração espacial (com quais países e localidades esses verbetes estão associados?) nem histórica / temporal (em quais anos, séculos ou épocas estão localizados os eventos associados com tais verbetes?).

Além disso, vários desses trabalhos relacionados não permitem uma visualização de uma composição personalizada pelo próprio usuário, isto é, onde o próprio usuário pode estabelecer um grupo de verbetes de seu interesse, não se limitando à exploração apenas da vizinhança de um nó num grafo. Ademais, os estudos citados não abordam uma característica que pode ser de grande importância na área de pesquisa do usuário: a análise do seu próprio histórico de pesquisas na Wikipédia e a criação de uma composição que integre essas informações com as demais características mencionadas. Os trabalhos relacionados abordam a Wikipédia, mas não o próprio histórico do usuário, deixando uma lacuna aberta. O presente trabalho não só permite o acesso por todos os canais anteriormente listados, como combina a exploração com o perfil do usuário, permitindo que ele explore suas próprias pesquisas, encontre tendências, reveja determinados aspectos, etc.

4 PROPOSTA

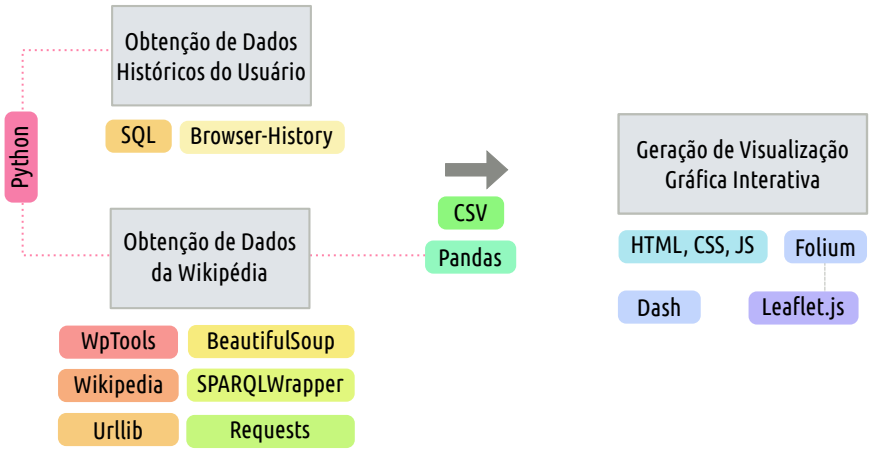
4.1 VISÃO GERAL

A ferramenta proposta no trabalho compõe-se de algumas funcionalidades, ou módulos, tendo sua implementação feita em *Python*, com o auxílio de tecnologias *web* (*HTML*, *CSS*, *Javascript*). Após ser iniciada, a aplicação solicita ao usuário um termo de seu interesse e o busca na Wikipédia, obtendo informações que serão utilizadas posteriormente para geração de visualizações gráficas.

Em termos de implementação, a proposta utiliza algumas técnicas, entre elas a de *Web Scraping*, todas suportadas pela linguagem *Python*. Para atingir os fins pretendidos, várias bibliotecas são utilizadas, como BeautifulSoup para manipulação do conteúdo *HTML*, Pandas para os dados, Requests para requisições *HTTP*, *Urllib* para manipulação de *URLs*, *CSV* para geração de arquivos nesse formato, entre outras. Para dar maior suporte as atividades de extração

de dados, também são utilizadas bibliotecas complementares, que encapsulam funcionalidades da própria *API* da Wikipédia, facilitando a busca pelas informações: *WpTools*[27] e *Wikipedia*[28].

Figura 2 – Visão geral da proposta



Fonte: Elaborado pelos autores.

Uma visão geral da aplicação proposta é apresentada na Figura 2. Seguindo a figura, a arquitetura da aplicação é constituída da seguinte maneira: (1) obtenção de dados através da Wikipédia/*Wikidata*, (2) obtenção de dados do histórico de navegação do usuário e (3) geração de gráficos interativos. Os passos 1 e 2 geram arquivos *CSVs* que serão usados para gerar as visualizações. Cada uma dessas partes é explorada mais detalhadamente a seguir.

4.2 OBTENÇÃO DE DADOS DA WIKIPÉDIA

Após identificado o termo a ser buscado, diversas rotinas de obtenção de dados são executadas. As informações necessárias nesse ponto da execução são extraídas através do código *HTML* do artigo correspondente ao termo buscado, tanto do corpo textual do artigo quanto das *Infoboxes*.

Tendo em vista que o usuário nem sempre conheça a forma exata de escrita do título de seu objeto de sua busca na Wikipédia, são realizadas ainda verificações e correções, caso necessárias, baseadas na entrada fornecida por ele. Essa entrada é utilizada para realizar diversas requisições *HTTP*, possibilitando a navegação pelo código *HTML* das várias páginas requisitadas, executando lógicas de busca a *tags* específicas para filtrar resultados, a fim de identificar corretamente, ou o mais próximo possível, o que se busca. Caso não seja possível encontrar nenhuma correspondência, o usuário é então informado.

[27] https://pypi.org/project/wptools/
[28] https://pypi.org/project/wikipedia/

Após a validação da entrada é mostrado um pequeno resumo do artigo encontrado pela ferramenta para dar *feedback* ao usuário e garantir que a busca seja a de interesse dele. Caso o usuário confirme as informações, a aplicação prossegue coletando demais informações.

Os dados escolhidos para serem coletados são de natureza biográfica básica sobre pessoas pesquisadas: Nome, Origem / Nacionalidade, Data de Nascimento e Falecimento, e Local de Nascimento e Falecimento. Esses dados são coletados mediante a confirmação do usuário sobre o artigo encontrado.

Como mencionado anteriormente, técnicas de *Web Scraping* são utilizadas para a obtenção de dados de sites da *web*, porém nem sempre os dados consultados dessa forma estão em formatos apropriados, dificultando ou mesmo tornando inviáveis determinadas extrações. Uma análise precisa da situação só pode ser feita caso a caso, considerando a natureza específica do site em questão e os dados envolvidos. Contudo, algumas plataformas oferecem meios alternativos de acesso aos dados que hospedam, como, por exemplo, por meio de uma *API*[29]. Dessa forma, o desenvolvedor pode interagir de forma mais direta com a *API* ou então usar alguma biblioteca de terceiros que já implemente as principais funcionalidades de consultas aos dados disponibilizados através da *API*. No caso da Wikipédia/*Wikidata*, o desenvolvedor pode realizar, se assim desejar, extração de vários tipos de dados parametrizados através de determinadas bibliotecas já disponíveis para várias linguagens. No presente trabalho, por se tratar de uma implementação com *Python*, optou-se pela utilização da biblioteca *SPARQLWrapper*, comumente adotada para tais operações[30].

Com a biblioteca *SPARQLWrapper* podem ser feitas consultas diretamente à *Wikidata* em linguagem *SPARQL*, tornando a busca por informações mais direta, embora seja necessário conhecimento técnico sobre a (não tão conhecida) linguagem por parte de quem adota esse meio. Na experiência de desenvolvimento deste trabalho, observou-se que a utilização da *SPARQLWrapper* para extração de dados possibilitou também a obtenção das coordenadas geográficas a partir do nome do país de origem da personalidade pesquisada, eliminando assim a necessidade de utilizar outras bibliotecas de consulta separadas com o objetivo específico (as quais resultavam em maior tempo de processamento e transmissão de dados). Essa abordagem contribuiu significativamente para aprimorar o desempenho da implementação, especialmente na geração das visualizações gráficas.

Através da biblioteca *SPARQLWrapper*, é possível estabelecer o endpoint *SPARQL* utilizando a URL do serviço que será alvo das consultas, como a *Wikidata*. Posteriormente, é necessário definir a consulta a ser executada, inserindo-a em formato de texto, seguindo a sintaxe específica da linguagem. Também é necessário especificar o formato de resposta desejado, como *JSON*. Por fim, a consulta é executada e os resultados são processados. Para facilitar a

[29] *Application Programming Interface*, interfaces que permitem a comunicação/integração de forma programada entre duas ou mais aplicações, sejam *web*, nativa, etc. Assim, sistemas e aplicações podem se comunicar e trocar recursos entre si.

[30] No momento de desenvolvimento deste trabalho, o projeto do *SPARQLWrapper* estava em primeiro lugar de buscas na plataforma Github para as palavras-chave "python"e "sparql", tendo o maior número de estrelas marcadas diante das demais opções.

criação das consultas, a plataforma da *Wikidata* já provê um serviço on-line para a criação e teste das mesmas, o chamado *Wikidata Query Service*[31], que possui recursos básicos de identação, *intellisense*, entre outros.

É importante ressaltar que não há uma garantia absoluta de que esses dados estarão sempre devidamente registrados na plataforma, nem mesmo de que eles estejam em estado atualizado e preciso. Essa condição depende principalmente das fontes de onde os dados são obtidos. É necessário considerar essas limitações da ferramenta devido à natureza parcialmente não estruturada dos dados na Wikipédia.

Após coletados, os dados são preprocessados a fim de retirar quaisquer outras informações desnecessárias. A implementação realizada aqui faz um tratamento posterior das datas e coordenadas geográficas, visando uma apresentação mais apropriada para o contexto. Com os dados devidamente coletados e limpos, eles são adicionados a um arquivo no formato *CSV*, utilizado pela ferramenta no módulo de geração de visualizações, conforme apresentado no fluxo da Figura 2.

4.3 OBTENÇÃO DE DADOS HISTÓRICOS DO USUÁRIO

Outra funcionalidade da ferramenta é a responsável pela obtenção dos dados de navegação do usuário. Inicialmente foram implementadas consultas manuais em *SQL* para isso. Porém, visando ampliar o acesso da ferramenta para várias combinações de outros navegadores e sistemas operacionais, optou-se por uma biblioteca em *Python* chamada *Browser-history*[32], que trabalha em uma amplitude maior de opções, realizando a detecção dos navegadores instalados na máquina do usuário, etc. Ela suporta uma variedade de navegadores populares, como *Chrome*, *Firefox*, *Safari* e *Microsoft Edge*. Isso garante que a ferramenta seja capaz de obter os dados de navegação independente do navegador utilizado pelo usuário. Após a coleta dos dados de navegação, são filtradas apenas as entradas referentes à Wikipédia. Essas entradas são então usadas para realizar buscas conforme descrito na seção anterior.

Uma vez que os resultados das buscas são obtidos e processados, o módulo finaliza sua execução salvando esses resultados em um arquivo *CSV*. A escolha do formato *CSV* para o armazenamento dos dados é comum devido à sua simplicidade e facilidade de manipulação. Esses dados serão então utilizados para a geração de gráficos, possibilitando uma visualização mais clara e intuitiva das informações coletadas.

4.4 GERAÇÃO DE GRÁFICOS INTERATIVOS

Com base nos dados adequadamente estruturados em um arquivo no formato *CSV*, o módulo mencionado nesta seção utiliza esse arquivo para gerar algumas visualizações para o usuário. As visualizações são construídas utilizando recursos das bibliotecas *Dash/Ploty*[33]

[31] https://query.wikidata.org/
[32] https://pypi.org/project/browser-history/
[33] https://dash.plotly.com/

e *Leaflet*[34], que aproveitam a flexibilidade da manipulação de dados da linguagem *Python*, combinada com a capacidade de gerar representações gráficas da linguagem *Javascript*.

Inicialmente, os dados são apresentados em um mapa usando marcadores no formato de pinos, cujas localizações aproximadas são baseadas em coordenadas geográficas obtidas nos processos descritos anteriormente. A precisão da localização pode variar conforme o termo pesquisado e as informações obtidas. Uma vez que o mapa é criado, os marcadores com as coordenadas são inseridos, e alguns ajustes são feitos para torná-los clicáveis, permitindo que as demais informações do arquivo *CSV* sejam inseridas e visualizadas por meio de um pop-up, semelhante a um cartão de informações. Essa personalização é viabilizada por meio da manipulação de códigos *HTML* inseridos nos marcadores.

O mapa também inclui um menu suspenso, onde é possível encontrar todos os registros pesquisados pela ferramenta, permitindo a busca textual e a seleção múltipla de entradas. Na parte inferior, há uma linha do tempo representada por um controle deslizante, apresentando os vários séculos relacionados às personalidades pesquisadas. É possível selecionar um século específico ou um intervalo deles. Todas as manipulações descritas são interativas, refletindo em tempo real na visualização do mapa. Por fim, são exibidas mais duas visualizações: um gráfico de barras e um gráfico de dispersão, oferecendo ao usuário alternativas para visualizar os dados.

Além disso, é importante destacar que todas as visualizações geradas são apresentadas em uma única página *web*, proporcionando uma experiência semelhante a um dashboard. Isso significa que o usuário pode acessar todas as visualizações de forma integrada e conveniente, permitindo uma análise abrangente dos dados em um único local. Essa abordagem centralizada facilita a compreensão das informações e oferece uma visão holística das diversas visualizações disponíveis, fornecendo uma interface intuitiva e amigável, permitindo ao usuário explorar e interagir com as visualizações de maneira eficiente.

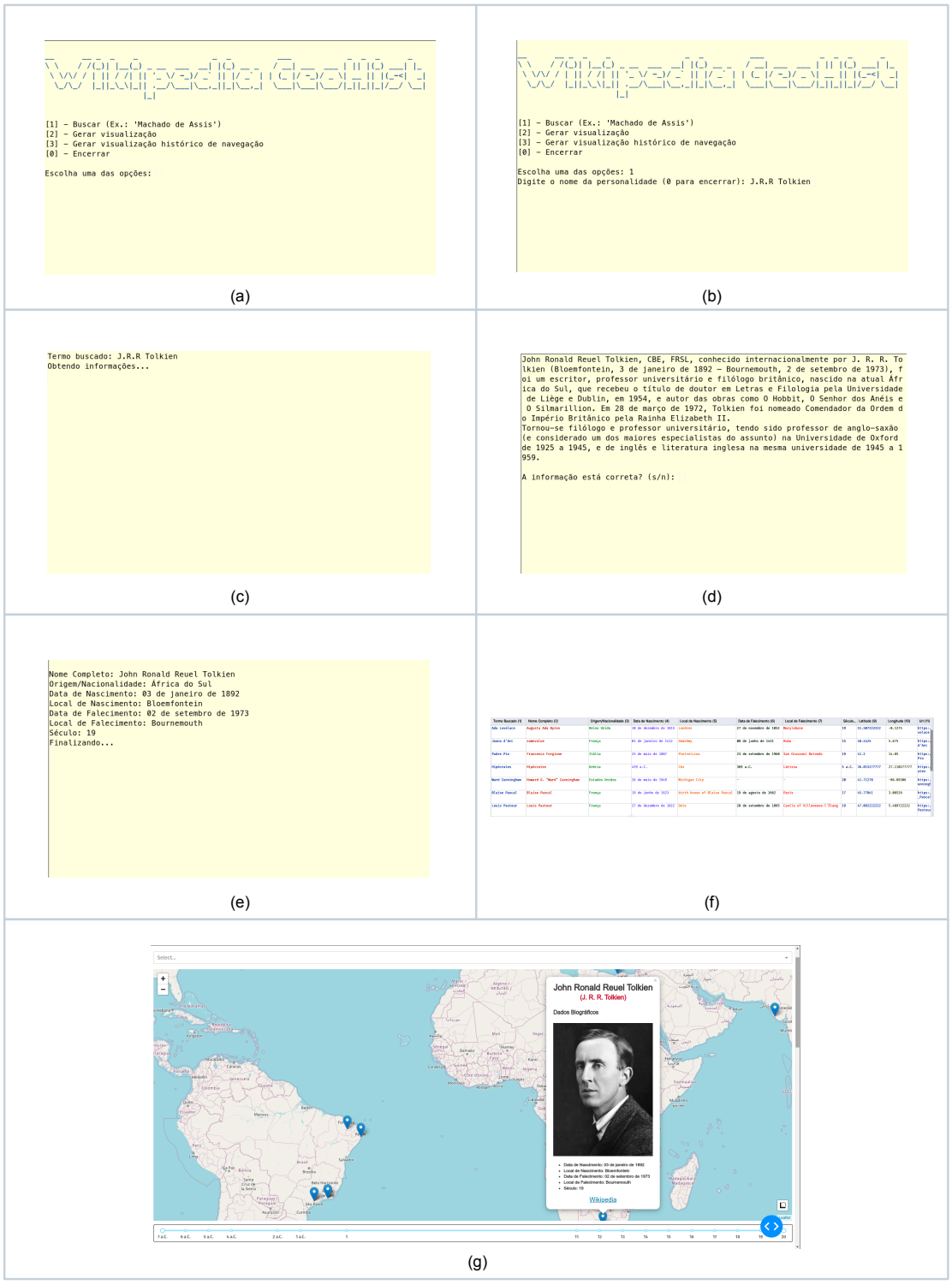
5 RESULTADOS

Nesta seção, são apresentados os resultados obtidos até o momento, ilustrando o fluxo de utilização da ferramenta, bem como o produto final gerado por ela, um mapa interativo contendo as buscas do usuário, dispostas geograficamente, e a demostração da obtenção dos dados a partir do histórico de navegação do mesmo.

Inicialmente a ferramenta foi desenvolvida como uma aplicação de terminal, também conhecida como aplicação de linha de comando, projetada para ser executado em um ambiente de terminal ou console de texto, onde o usuário interage com o programa por meio de comandos de texto inseridos em um prompt. O primeiro contato do usuário com a ferramenta se dá por meio de um menu, conforme mostrado na Figura 3a. Dentre as opções disponíveis, o usuário pode (1) fazer a sua busca, (2) gerar suas visualizações baseadas nas buscas feitas na opção anterior, (3) gerar visualizações baseadas no histórico de navegação e (0) encerrar a aplicação. Escolhendo a primeira opção, é solicitado o termo a ser buscado, mostrado na Figura 3b.

[34] https://leafletjs.com/

Figura 3 – Telas do Sistema



Fonte: Elaborado pelos autores.

Ao escolher a primeira opção e inserir o termo de busca desejado, a ferramenta utiliza essa entrada para realizar possíveis correções, garantindo uma busca bem-sucedida. Durante esse processo, o usuário é informado de maneira transparente e deve aguardar alguns segundos,

conforme indicado na Figura 3c. Em seguida é apresentado um trecho inicial do artigo encontrado, solicitando a confirmação das informações, como na Figura 3d. Caso o resultado não seja satisfatório, ou o usuário deseje refazer sua buscar, ele deve responder “Não” (n), sendo assim solicitado novamente um termo a ser buscado. Após 3 negativas, a aplicação se encerrar, sugerindo ao usuário o refinamento da sua busca. Em caso positivo, a ferramenta concluirá o processo de extração das informações, e em alguns segundos, será mostrado um resumo do que foi obtido, conforme a Figura 3e.

Ao fim do ciclo de buscas, a aplicação retorna ao menu inicial e as informações obtidas são armazenadas em um arquivo no formato *CSV*, de maneira estruturada, como visto na Figura 3f, para posterior uso no módulo de geração de visualizações da ferramenta.

Com o usuário tendo feito todas as suas buscas, é possível gerar sua visualização, escolhendo a opção 2 do menu. Utilizando o arquivo *CSV* gerado durante o processo de extração de informações, a ferramenta constrói um mapa interativo contendo as buscas realizadas, localizadoas geograficamente por meio de pinos, que ao serem clicados, exibem *cards* de informações, permitindo também a livre navegação pelo mapa. Na parte superior do mapa, há um menu suspenso com todos os resultados buscados, que podem ser selecionados individualmente ou em grupo, alterando a visão do mapa. Na parte inferior encontra-se um menu deslizante, contendo os séculos de todos os resultados, os quais podem ser selecionado de forma única, ou intervalo deles. O mapa é gerado e automaticamente aberto no navegador do usuário para visualização, como visto na Figura 3g. Após todo o ciclo de execução, os dados coletados ficam gravados em arquivo no formato *CSV*, possibilitando consultas posteriores ou compartilhamento de dados caso o usuário deseje.

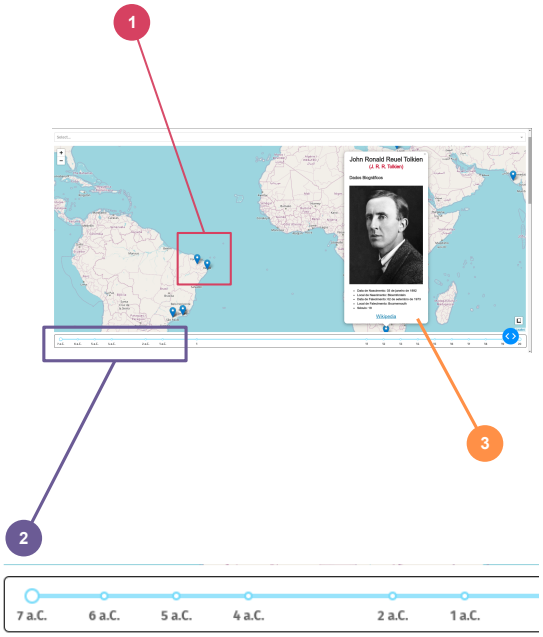
Entre os detalhes apresentados na visualização do mapa (Figura 4), encontram-se (1) marcadores de outras buscas realizadas; (2) linha de tempo, mostrando os séculos com dados encontrados; (3) *card* de informações que aparece ao clicar num marcador (pop-up). Apenas os marcadores que estão dentro da faixa de tempo selecionada são exibidos, permitindo que o usuário realize filtro.

O *card* (Figura 5), por sua vez, apresenta (1) a identificação do registro; (2) uma imagem baixada na hora a partir da Wikipédia; (3) dados puxados da Wikipédia; e (4) um *link* para o artigo original da Wikipédia sobre o registro em questão.

Quanto às visualizações do histórico do usuário, a ferramenta possui um módulo responsável pela extração de seus dados. Ao executar a aplicação e escolher a opção (3) do menu, os registros pertinentes à Wikipédia são selecionados, permitindo a realização de buscas com resultados específicos nessa plataforma. Após a coleta dos dados, os processos de armazenamento em formato *CSV* e a geração das visualizações são os mesmos descritos anteriormente. O resultado gráfico gerado pela ferramenta para um determinado histórico de usuário é apresentado na Figura 6. Nesse gráfico, cada registro buscado na Wikipédia, ao longo do uso do navegador por parte do usuário, foi levado em consideração e então classificado quanto ao país e época, permitindo o usuário uma inspeção sobre seu próprio histórico de buscas.

Também é possível obter uma visualização do mapa com marcadores referentes aos

Figura 4 – Detalhes da visualização do mapa

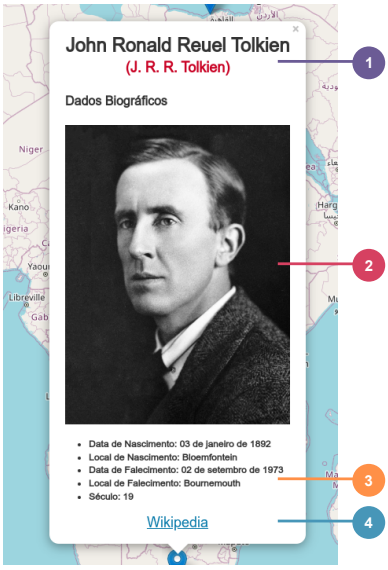


Fonte: Elaborado pelos autores.

diversos registros já pesquisados pelo usuário, conforme ilustrado na Figura 7.

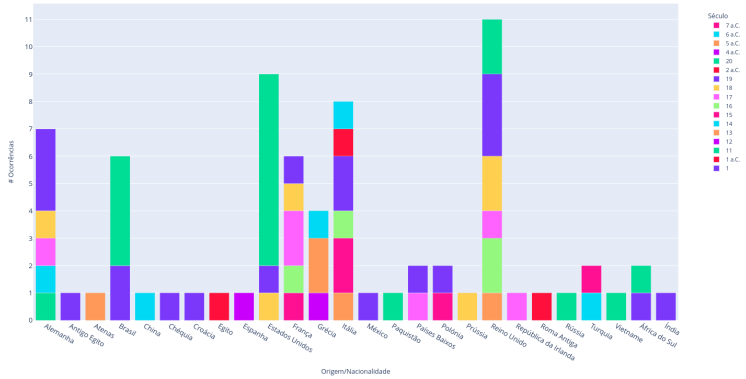
Como mencionado anteriormente, esse processo de extração e análise do histórico de navegação proporciona uma compreensão mais aprofundada do horizonte literário do usuário

Figura 5 – Detalhes do card



Fonte: Elaborado pelos autores.

Figura 6 – Gráfico para o histórico de buscas de um usuário ao longo do tempo, mostrando as tendências e os interesses do usuário em diferentes períodos.



Fonte: Elaborado pelos autores.

Figura 7 – Mapa com marcadores mostrando a localização dos diferentes resultados de busca.

Fonte: Elaborado pelos autores.

na plataforma, permitindo *insights* que podem ser valiosos. É possível empreender uma análise de padrões de pesquisa, preferências de conteúdo e tendências de interesse (ver, por exemplo, um determinado perfil de usuário quanto às suas buscas na Figura 8, permitindo ao usuário uma melhor percepção sobre algum potencial tipo de *bias* em suas fontes, etc).

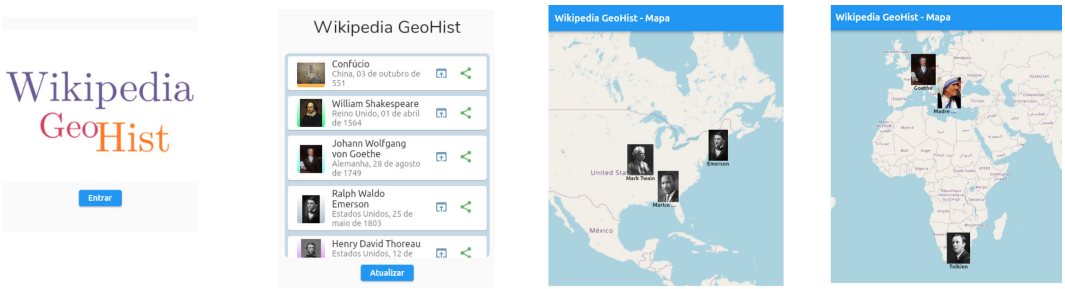
Além da implementação principal, uma implementação adicional da visualização dos registros foi conduzida em *Flutter*, para facilitar a exploração de tais resultados em plataformas móveis, conforme apresentado na Figura 9. *Flutter* é um *framework* de desenvolvimento com foco multiplataforma em dispositivos móveis, como *Android* e *iOS*. Criado pela *Google*, é bastante utilizado no mercado e, mais recentemente, permitiu a criação de aplicações para *desktop (Linux, Windows e macOS)*. Seus principais benefícios são a versatilidade, menor curva de aprendizado e agilidade. Essa implementação trabalha com outra forma / dimensão de tela,

Figura 8 – Sugestão de visualização gráfica criada partir do histórico de consultas

Fonte: Elaborado pelos autores.

além de focar também em operações típicas no segmento móvel, como compartilhar itens da listagem. A exibição de imagens para as buscas realizadas ocorre em tempo real, sob demanda conforme o usuário usa o programa. A implementação para dispositivos móveis está em estágio inicial de desenvolvimento e ainda deve passar por mais testes antes de ser disponibilizada, embora os experimentos iniciais tenham sido promissores.

Figura 9 – Telas do protótipo da aplicação para dispositivos móveis.



Fonte: Elaborado pelos autores.

6 CONSIDERAÇÕES FINAIS

No contexto de um mundo cada vez mais interconectado, com acesso contínuo a diversas fontes de informação, a Wikipédia tem se destacado como uma plataforma pública e gratuita de informações enciclopédicas. Nesse cenário, surge a necessidade crescente de otimizar o tempo e apresentar informações de maneira mais didática. Este estudo apresentou, como proposta principal, uma ferramenta que permite a geração de visualizações gráficas rápidas de termos de interesse, tanto por buscas personalizadas como pelo histórico de navegação do usuário na Wikipédia. Tal recurso possibilita a realização de análises históricas e geográficas de forma ágil, bem como dá subsídios para uma análise das próprias tendências nas buscas e pesquisas realizadas, sendo relevante para pesquisadores, estudantes e entusiastas, permitindo enriquecer suas experiências de pesquisa. Também apresentou uma compilação de ferramentas relacionadas à Wikipédia, inseridas no contexto de exploração e extração de dados.

A proposta de ferramenta abrange possibilidades de utilização tanto por desenvolvedores como por usuários finais. Além disso, a ferramenta é disponibilizada como um software de código aberto, sendo acessível para uso e análise pela comunidade por meio de um repositório oficial. A disseminação da informação e de mecanismos que facilitam sua compreensão, tanto em questões visuais e didáticas, facilitando conexões, como em questões de autoanálise da pesquisa, contribuem para o conhecimento e aprendizado.

Para trabalhos futuros, é válido explorar o potencial de utilização da DBpédia e o aprofundamento na *Wikidata*, incluindo seus tipos de dados, estruturação e linguagem de consulta específica. Além disso, é possível considerar o uso de visualizações gráficas 3D, particularmente através de *frameworks* como *ThreeJS*, para abrir outras perspectivas visuais. Por fim, pode ser valioso ainda desenvolver uma extensão ou *plug-in* para navegadores populares, como *Chrome* ou *Firefox*, a fim de aprimorar a interface gráfica durante a navegação na Wikipédia, integrando com algumas das várias funcionalidades descritas ao longo do trabalho. Essas melhorias visam proporcionar uma experiência mais enriquecedora para os usuários.

REFERÊNCIAS

AUER, S. et al. Dbpedia: A nucleus for a web of open data. In: **The semantic web**. [S.l.]: Springer, 2007. p. 722–735.

BHOSALE, S.; PATIL, M. T.; PATIL, M. P. Sqlite: Light database system. **Int. J. Comput. Sci. Mob. Comput**, v. 44, n. 4, p. 882–885, 2015.

BIUK-AGHAI, R. P.; PANG, C.-I.; SI, Y.-W. Visualizing large-scale human collaboration in wikipedia. **Future Generation Computer Systems**, Elsevier, v. 31, p. 120–133, 2014.

CHAN, B. et al. Vispedia: Interactive visual exploration of wikipedia data via search-based integration. **IEEE Transactions on Visualization and Computer Graphics**, IEEE, v. 14, n. 6, p. 1213–1220, 2008.

CHASIN, R. Event and temporal information extraction towards timelines of wikipedia articles. **Simile**, p. 1–9, 2010.

HIENERT, D.; LUCIANO, F. Extraction of historical events from wikipedia. In: SPRINGER. **Extended Semantic Web Conference**. [S.l.], 2012. p. 16–28.

JAMALI, H. R.; ASADI, S. Google and the scholar: the role of google in scientists' information-seeking behaviour. **Online information review**, Emerald Group Publishing Limited, 2010.

JAMIL, G. L.; NEVES, J. T. de R. A era da informação: considerações sobre o desenvolvimento das tecnologias da informação. **Perspectivas em ciência da informação**, v. 5, n. 1, 2000.

KERN, V. M. A wikipédia como fonte de informação de referência: avaliação e perspectivas. **Perspectivas em ciência da informacão**, SciELO Brasil, v. 23, p. 120–143, 2018.

MCBRIDE, B. The resource description framework (rdf) and its vocabulary description language rdfs. In: **Handbook on ontologies**. [S.l.]: Springer, 2004. p. 51–65.

MILNE, D.; WITTEN, I. H. An open-source toolkit for mining wikipedia. **Artificial Intelligence**, Elsevier, v. 194, p. 222–239, 2013.

OLTEANU, D. et al. Xpath: looking forward. In: SPRINGER. **International Conference on Extending Database Technology**. [S.l.], 2002. p. 109–127.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of sparql. In: SPRINGER. **International semantic web conference**. [S.l.], 2006. p. 30–43.

RAMAN, M. Wiki technology as a "free" collaborative tool within an organizational setting. **EDPACS**, Taylor & Francis, v. 42, n. 3, p. 1–10, 2010.

REBELE, T. et al. Yago: A multilingual knowledge base from wikipedia, wordnet, and geonames. In: SPRINGER. **International semantic web conference**. [S.l.], 2016. p. 177–185.

RUSSO, I.; CASELLI, T.; MONACHINI, M. Extracting and visualising biographical events from wikipedia. In: **BD**. [S.l.: s.n.], 2015. p. 111–115.

SANGER, L. The early history of nupedia and wikipedia: a memoir. **Open sources**, O'Reilly Media Sebastopol, CA, v. 2, p. 307–38, 2005.

SIPOŠ, R. et al. Historyviz–visualizing events and relations extracted from wikipedia. In: SPRINGER. **European Semantic Web Conference**. [S.l.], 2009. p. 903–907.

UTOMO, M. S. Web scraping pada situs wikipedia menggunakan metode ekspresi regular. **Dinamik**, v. 18, n. 2, 2013.

WALLACE, D. P.; FLEET, C. V. From the editors: The democratization of information? wikipedia as a reference resource. **Reference & User Services Quarterly**, JSTOR, p. 100–103, 2005.

WANG, Y. et al. Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In: **Proceedings of the 13th International Conference on Extending Database Technology**. [S.l.: s.n.], 2010. p. 697–700.