INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
DE MINAS GERAIS (IFMG)
CAMPUS BAMBUÍ
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Gustavo Silveira Dias

REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MOTORES DE BUSCA NA RECUPERAÇÃO DE INFORMAÇÕES

Bambuí – MG
2026

GUSTAVO SILVEIRA DIAS

REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MOTORES DE BUSCA NA RECUPERAÇÃO DE INFORMAÇÕES

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Felipe Lopes de Melo Faria

Bambuí – MG
2026

Catalogação na Fonte Biblioteca IFMG - *Campus* Bambuí

D541r Dias, Gustavo Silveira.

Revisão sistemática da literatura sobre motores de busca na recuperação de informações [manuscrito] / Gustavo Silveira Dias. – 2026.

63 f. : il. ; color.

Orientador: Felipe Lopes de Melo Faria.
Trabalho de Conclusão de Curso (Bacharelado em Engenharia de Computação) – Instituto Federal de Minas Gerais. Campus Bambuí, 2026.

1. Recuperação de informação. 2. Motores de busca. 3. Modelos de busca. I. Faria, Felipe Lopes de Melo. II. Instituto Federal de Minas Gerais – Campus Bambuí. III. Título

CDD 025.04

Catalogação: João Batista Rodrigues - CRB-6/2022

Gustavo Silveira Dias

REVISÃO SISTEMÁTICA DA LITERATURA SOBRE MOTORES DE BUSCA NA RECUPERAÇÃO DE INFORMAÇÕES

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais (IFMG) – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Aprovado em 13 de Maio de 2026 pela banca examinadora:

Felipe Lopes de Melo Faria – IFMG – *Campus* Bambuí – (Orientador)
Bruno Alberto Soares Oliveira – IFMG - Campus Bambuí
Natalia Camillo do Carmo – IFMG - Campus Bambuí

Para todos aqueles que, assim como eu, buscam desbravar o mundo dos Motores de Busca.

AGRADECIMENTOS

Agradeço primeiramente, a Deus por me conceder saúde, força e perseverança para superar os desafios ao longo dessa trajetória.

Agradeço aos meus pais e familiares o apoio, incentivo e compreensão em todos os momentos, que foram fundamentais para a conclusão desse curso.

Em especial, agradeço ao meu orientador a orientação, paciência, disponibilidade e valiosas contribuições para o desenvolvimento deste trabalho.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho de Conclusão de Curso.

“O maior sofrimento é sentir saudades de algo que talvez nunca mais volte.”
Arthur Schopenhauer

RESUMO

O crescimento exponencial da informação digital tornou os motores de busca ferramentas essenciais para a recuperação de dados em diversos contextos acadêmicos e profissionais. Embora já existam estudos relevantes sobre o tema, a literatura ainda necessita avançar gradativamente na sistematização dos principais motores de busca e na avaliação de sua eficácia na recuperação de informações relevantes. Este trabalho teve como propósito realizar uma revisão sistemática da literatura sobre motores de busca aplicados à recuperação de informações, analisando seus pontos fortes, limitações e áreas de aplicação. Pretende-se oferecer uma categorização dos estudos analisados que auxilie pesquisadores a entender o estado da arte atual. A metodologia adotada baseou-se em uma revisão sistemática da literatura, permitindo identificar, analisar e sintetizar estudos acadêmicos relevantes sobre motores de busca aplicados à recuperação de informação. Definiram-se critérios de relevância para seleção das fontes, além da escolha de palavras-chave, bases de dados e delimitação do corpus da pesquisa. A aplicação desses critérios permitiu uma análise crítica dos materiais selecionados, visando à organização e discussão dos resultados de forma estruturada. Espera-se que o estudo contribua para uma compreensão aprofundada dos motores de busca como mediadores do acesso à informação, destacando seus impactos na qualidade, relevância e imparcialidade dos resultados. Ao identificar as principais abordagens e algoritmos utilizados, bem como mapear tendências e lacunas na literatura, pretende-se auxiliar futuros pesquisadores no desenvolvimento de investigações mais eficazes no campo da recuperação de informação. Dessa forma, o trabalho busca integrar conhecimentos da Ciência da Computação e da Ciência da Informação, promovendo uma visão mais crítica e abrangente sobre o papel dessas ferramentas no ambiente digital contemporâneo.

Palavras-chave: Recuperação de Informação. Motores de busca. Modelos de Busca.

ABSTRACT

Exponential growth of digital information has made search engines essential tools for data retrieval in various academic and professional contexts. Although there are already relevant studies on the subject, the literature still needs to gradually advance in the systematization of the main search engines and in the evaluation of their effectiveness in retrieving relevant information. This work aims to conduct a systematic review of the literature on search engines applied to information retrieval, analyzing their strengths, limitations and areas of application. The aim is to provide a categorization of the analyzed studies that will help researchers understand the current state of the art. The methodology adopted is based on a systematic review of the literature, allowing the identification, analysis and synthesis of relevant academic studies on search engines applied to information retrieval. Relevance criteria will be defined for the selection of sources, in addition to the choice of keywords, databases and delimitation of the research corpus. The application of these criteria will allow a critical analysis of the selected materials, aiming at the organization and discussion of the results in a structured manner. The study is expected to contribute to a deeper understanding of search engines as mediators of information access, highlighting their impacts on the quality, relevance and impartiality of results. By identifying the main approaches and algorithms used, as well as mapping trends and gaps in the literature, it is intended to assist future researchers in developing more effective investigations in the field of information retrieval. In this way, the work seeks to integrate knowledge from Computer Science and Information Science, promoting a more critical and comprehensive view of the role of these tools in the contemporary digital environment.

Keywords: Information Retrieval. Search Engines. Search Models.

LISTA DE FIGURAS

Figura 1 - Fluxograma da RSL 27
Figura 2 - Funcionamento Scrum 28
Figura 3 - Sistema Archie 31
Figura 4 - Sistema Gopher 33
Figura 5 - Arquitetura do Web Crawler 42
Figura 6 - Arquitetura de um mecanismo de busca 47

LISTA DE TABELAS

Tabela 1 - Backlog do projeto 28
Tabela 2 - *Sprint* 1 28
Tabela 3 - *Sprint* 2 28
Tabela 4 - *Sprint* 2 29
Tabela 5 - *Sprint* 3 29
Tabela 6 - *Sprint* 3 29
Tabela 7 - Itens do Menu 33
Tabela 8 - Modos de Consulta do AltaVista 43
Tabela 9 - Etapas do Processo de Indexação do AltaVista 44
Tabela 10 -Operadores de Consulta do AltaVista 45

LISTA DE SIGLAS

ABNT – Associação Brasileira de Normas Técnicas

BERT – Bidirectional Encoder Representations for Transformers

CC – Ciência da Computação

CERN – Conseil Européen pour la Recherche Nucléaire

CI – Ciência da Informação

DEC – Digital Equipment Corporation

FTP – File Transfer Protocol

GCC – GNU Compiler Collection

HTML – HyperText Markup Language

HTTP – HyperText Transfer Protocol

IA – Inteligência Artificial

IDF – Inverso da Frequência do Documento

IETF – Internet Engineering Task Force

IFMG – Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais

IFSP – Instituto Federal de São Paulo

IP – Internet Protocol

MLM – Masked Language Modeling

MUM – Multitask Unified Model

NSP – Next Sentence Prediction

PLN – Processamento de Linguagem Natural

RFC – Request For Comments

RI – Recuperação de Informação

RSL – Revisão Sistemática da Literatura

SI – Sistema de Informação

TCC – Trabalho de Conclusão de Curso

TCP – Transmission Control Protocol

TF – Frequência do Termo

TI – Tecnologia da Informação

UCL – University College London

UFRJ – Universidade Federal do Rio de Janeiro

UFSC – Universidade Federal de Santa Catarina

URL – Uniform Resource Locator

USP – Universidade de São Paulo

WPR – Weighted PageRank

WWW – World Wide Web

SUMÁRIO

1 INTRODUÇÃO 15

1.1 Objetivos 16

1.1.1 Objetivo geral 16

1.1.2 Objetivos específicos 16

1.2 Justificativa 17

1.3 Estrutura do trabalho 17

2 REVISÃO BIBLIOGRÁFICA 18

2.1 Fundamentação teórica 18

2.1.1 Modelos Quantitativos 18

2.1.1.1 Modelo Vetorial 18

2.1.1.2 Modelo Booleano 20

2.1.1.3 Modelo Fuzzy 20

2.1.2 Modelos Dinâmicos 21

2.1.2.1 Sistemas especialistas 21

2.1.2.2 Redes neurais 22

2.1.2.3 Algoritmos genéticos 22

2.2 Estado da arte 22

3 METODOLOGIA 24

3.1 Classificação da pesquisa 24

3.2 Protocolo de revisão sistemática da literatura 24

3.3 Scrum 27

3.4 Materiais e tecnologias 29

4 RESULTADOS 30

4.1 Análise do Modelo de Funcionamento do Motor Archie 30

4.1.1 Aplicações Práticas do Archie 31

4.2 Análise do Modelo de Busca Gopher 32

4.2.1 Aplicações Práticas do Gopher 34

4.2.2 Vantagens e Desvantagens 35

4.2.2.1 Vantagens 35

4.2.2.2 Desvantagens 35

4.3 O Surgimento do Veronica no Gopherspace 36

4.3.1 O Processo de Indexação Automatizada 36

4.3.2 Limitações da Indexação e Atualização 37

4.3.3 Vantagens e Desvantagens 37

4.3.3.1 Vantagens do Veronica na Época 38

4.3.3.2 Desvantagens do Veronica na Época 38

4.4 Jughead 38

4.4.1 Aplicações Práticas do Jughead 39

4.4.2 Vantagens e Desvantagens 40

4.4.2.1 Vantagens do Jughead na Época 40

4.4.2.2 Desvantagens do Jughead na Época 40

4.5 Surgimento e Mecanismo de Funcionamento do Web Crawler 41

4.6 Funcionamento do Motor de Busca AltaVista 42

4.6.1 Arquitetura Funcional do AltaVista 43

4.6.2 Rastreamento e Coleta de Páginas Web 43

4.6.3 Indexação e Organização da Informação 44

4.6.4 Consultas e Recuperação dos Resultados 44

4.6.4.1 Modos de Operação 45

4.6.4.2 Operadores de Consulta e Lógica Booleana 45

4.6.4.3 Ordenação e Relevância 45

4.6.5 Inovações e Contribuições Tecnológicas 45

4.6.6 Cadê? 46

4.7 Google: A Revolução na Recuperação de Informação 46

4.7.1 Arquitetura geral de um mecanismo de busca moderno 46

4.7.2 O algoritmo PageRank: ideia e formulação 47

4.7.3 Como o PageRank se integra ao processo de ranking 48

4.7.4 Evolução dos algoritmos de ranking (principais marcos) 49

4.7.5 Desafios, controvérsias e considerações éticas 50

4.8 Redes Sociais como Motores de Busca ? 50

4.8.1 Mudança de paradigma no comportamento de busca 51

4.8.2 Características que permitem o funcionamento como motor de busca 51

4.8.3 Potencialidades e desafios 52

4.8.4 Implicações para a Ciência da Informação e para práticas de busca . 52

5 CONCLUSÃO 53

REFERÊNCIAS 55

1 INTRODUÇÃO

Desde os primórdios da Internet, existe a necessidade de desenvolver mecanismos capazes de localizar informações de maneira rápida e eficiente. Entre as primeiras iniciativas voltadas a esse propósito destacam-se o Archie, utilizado para localizar arquivos em repositórios *File Transfer Protocol* (FTP), além do Veronica e do Jughead, responsáveis pela busca em servidores Gopher (Cendón, 2001). Contudo, foi com o surgimento da *World Wide Web* (WWW) e a expansão acelerada da quantidade de conteúdos digitais publicados que a busca por informação passou a representar um desafio ainda maior. Atualmente, estima-se a existência de bilhões de páginas em formato *HyperText Markup Language* (HTML), tornando indispensável o uso de ferramentas especializadas para recuperar informações relevantes em meio ao grande volume de dados disponíveis (CARDILLO, 2025).

Nesse cenário, os motores de busca consolidaram-se como instrumentos fundamentais para a organização e recuperação da informação na internet. Esses sistemas são responsáveis por percorrer páginas *web*, indexar conteúdos e retornar resultados de acordo com as consultas realizadas pelos usuários (Ferneda, 2003). Ao efetuar uma pesquisa, o usuário recebe uma lista ordenada de documentos considerados mais relevantes para sua necessidade informacional, facilitando o acesso rápido e eficiente ao conteúdo desejado (DOMINGUES, 2019).

Para proporcionar resultados cada vez mais precisos, os motores de busca utilizam diferentes técnicas de Recuperação da Informação (RI). Entre os modelos mais relevantes destaca-se o modelo vetorial, amplamente empregado em sistemas de busca devido à sua capacidade de representar documentos e consultas como vetores em um espaço n-dimensional (Souza, 2006). Nesse modelo, os termos recebem pesos conforme sua importância relativa dentro dos documentos, permitindo calcular o grau de similaridade entre uma consulta e os conteúdos indexados. Diferentemente do modelo booleano, que depende de correspondências exatas entre palavras, o modelo vetorial possibilita identificar documentos parcialmente relacionados à consulta, ampliando a relevância dos resultados retornados (DOMINGUES, 2019).

A comparação entre documentos e consultas ocorre por meio do cálculo do cosseno do ângulo formado entre os vetores representativos. Quanto menor o ângulo entre eles, maior será o grau de similaridade identificado pelo sistema (BAEZA-YATES *et al.*, 1999) citado por (DOMINGUES, 2019). Esse mecanismo permite estabelecer um ranqueamento mais eficiente dos resultados, contribuindo para uma experiência de busca mais alinhada à intenção do usuário.

O estudo desses mecanismos está diretamente relacionado à Ciência da Informação (CI), área que investiga os processos de produção, organização, representação, recuperação e uso da informação em diferentes contextos sociais e tecno-

lógicos (Araújo, 2018). Segundo (Ferneda, 2003), a RI configura-se como um dos principais ramos da CI, sendo responsável pelo desenvolvimento de métodos e sistemas capazes de localizar informações relevantes em grandes volumes de dados digitais.

Com o crescimento contínuo das informações disponíveis na internet, os sistemas de RI passaram por constantes transformações. De acordo com (Souza, 2006), esses sistemas deixaram de se apoiar exclusivamente em palavras-chave e passaram a incorporar aspectos mais complexos, como contexto, relevância e necessidades específicas dos usuários. Dessa forma, observa-se uma tendência no desenvolvimento de métodos cada vez mais inteligentes e eficientes, capazes de compreender melhor o significado das consultas e oferecer resultados mais personalizados e precisos. Essa evolução busca atender à crescente demanda por mecanismos de busca mais eficazes em um ambiente digital caracterizado pela intensa produção e circulação de informações.

1.1 Objetivos

Esta seção apresenta os objetivos gerais que orientaram o desenvolvimento deste estudo, bem como os objetivos específicos, que detalharam as etapas necessárias para alcançá-los.

1.1.1 Objetivo geral

Realizar uma Revisão Sistemática da Literatura Sobre Motores de Busca na Recuperação de Informação.

1.1.2 Objetivos específicos

Os objetivos específicos detalham etapas intermediárias para se alcançar o objetivo geral.

a) Realizar levantamento de artigos relacionados a Motores de Busca no contexto de Recuperação de Informação;
b) Analisar os resultados no levantamento de informações, elecando características da técnica estudada;
c) Analisar exemplos de aplicações de motores de busca;
d) Disponibilizar uma linha histórica dos motores de busca.

1.2 Justificativa

Segundo (Ferneda, 2003), a *web* representa a face hipertextual da Internet, consolidando-se como uma das principais fontes de informação nas mais diversas áreas do conhecimento. Com aproximadamente 1 bilhão de sites disponíveis, o volume de conteúdos publicados cresce de forma exponencial, influenciando diretamente a maneira como a sociedade busca, filtra e consome informações no ambiente digital (CARDILLO, 2025).

Nesse contexto de expansão informacional, os motores de busca assumem um papel central, pois são responsáveis por localizar, organizar e disponibilizar conteúdos relevantes aos usuários (Ferneda, 2003). Mais do que simples ferramentas computacionais, esses sistemas tornaram-se mediadores essenciais do acesso ao conhecimento, influenciando a forma como as informações são encontradas e utilizadas. Apesar de sua relevância no cenário contemporâneo, não se identificam trabalhos que apresentem, de maneira abrangente, uma linha histórica da evolução dos motores de busca. Diante disso, torna-se importante investigar esses sistemas tanto sob a perspectiva tecnológica quanto como instrumentos fundamentais de mediação da informação e do conhecimento.

O estudo insere-se nesse contexto ao buscar compreender e aprimorar os processos de recuperação de informação, especialmente em ambientes digitais dinâmicos e com alto volume de dados. Desse modo, academicamente, a pesquisa contribui para preencher lacunas na literatura ao integrar fundamentos da Ciência da Computação (CC) e da CI, promovendo, assim, uma visão mais abrangente e crítica sobre o tema.

1.3 Estrutura do trabalho

O presente trabalho está estruturado da seguinte maneira: o Capítulo (2) apresenta a revisão bibliográfica, abordando os principais conceitos e o estado da arte na área estudada. Posteriormente, dentro desse mesmo capítulo, é dedicada uma seção à fundamentação teórica, com ênfase nos modelos quantitativos e dinâmicos de RI, detalhando as técnicas e metodologias associadas a cada abordagem.

O capítulo de Metodologia (3) descreve os procedimentos adotados para o desenvolvimento da pesquisa, incluindo os métodos, técnicas e ferramentas utilizadas na coleta, processamento e análise dos dados. Em seguida, o capítulo de Resultados (4) apresenta os principais achados obtidos no presente estudo,discutindo-os à luz dos objetivos propostos e da literatura revisada. Por fim, o capítulo de Conclusão (5) sintetiza as contribuições do trabalho, destacando suas implicações, limitações e possíveis direções para pesquisas futuras.

2 REVISÃO BIBLIOGRÁFICA

A revisão bibliográfica deve analisar criticamente o estado da arte sobre o tema estudado, identificando principais contribuições, avanços recentes e lacunas ainda existentes na literatura.

2.1 Fundamentação teórica

A fundamentação teórica tem por objetivo apresentar os principais conceitos, modelos e abordagens relacionados ao tema do trabalho, servindo de base para a compreensão das propostas e análises desenvolvidas. Nesta seção, serão discutidos os métodos clássicos de RI.

2.1.1 Modelos Quantitativos

Os modelos quantitativos de RI baseiam-se em disciplinas como lógica, estatística e teoria dos conjuntos (Ferneda, 2003). Segundo (Robertson, 1977), seu predomínio se deve à necessidade de análise formal e à formulação explícita das hipóteses envolvidas.

Esses modelos representam documentos por meio de termos de indexação, ou seja, palavras que expressam conceitos presentes no conteúdo (Ferneda, 2003). No entanto, nem todos os termos têm o mesmo valor representativo: alguns são centrais, enquanto outros são periféricos.

A importância de um termo pode ser medida com base em sua frequência no corpus, de modo que termos muito comuns têm pouca utilidade, enquanto termos raros podem ser mais informativos para filtrar documentos relevantes (DOMINGUES, 2019). Assim, os modelos quantitativos buscam atribuir diferentes pesos aos termos conforme sua capacidade de representar com precisão o conteúdo dos documentos.

2.1.1.1 Modelo Vetorial

O modelo vetorial, também conhecido como modelo espaço vetorial, é um dos principais modelos utilizados na RI. Ele foi proposto por Gerard Salton na década de 1970 e se tornou uma referência na área devido à sua eficácia na representação e comparação de documentos. Baseia-se na ideia de representar documentos e consultas como vetores em um espaço multidimensional (Souza, 2006). Cada dimensão desse espaço corresponde a um termo (palavra-chave) distinto presente no vocabulário do corpus (DOMINGUES, 2019). A motivação principal por trás do modelo é permitir uma medição quantitativa da similaridade entre a consulta e a expressão de

busca (Ferneda, 2003).

Em termos práticos, um documento é representado como um vetor de pesos, sendo que cada peso reflete a relevância de um termo específico dentro do corpus (Ferneda, 2003). Esses pesos, geralmente, são calculados utilizando-se a métrica *Term Frequency-Inverse Document Frequency* (TF-IDF) (SOUZA *et al.*, 2017). A frequência do termo (TF) representa quantas vezes a palavra aparece no documento, enquanto o inverso da frequência nos documentos (IDF) penaliza termos comuns, priorizando palavras mais específicas e, portanto, mais discriminativas (Ferneda, 2003). Essa ponderação contribui para que a representação vetorial capture a importância real de cada termo no contexto do corpus.

O cálculo da frequência do termo pode ser definido como:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t'} f_{t',d}} \tag{2.1}$$

em que $f_{t,d}$ representa o número de ocorrências do termo t no documento d, e o denominador corresponde ao total de termos no documento. Já o inverso da frequência de documentos é calculado por:

$$IDF(t) = \log\left(\frac{N}{df_t}\right) \tag{2.2}$$

em que N é o número total de documentos na coleção, e df_t corresponde ao número de documentos que contêm o termo t. Dessa forma, o peso final de um termo em um documento é dado por:

$$w_{t,d} = TF(t,d) \times IDF(t) \tag{2.3}$$

Após o cálculo dos pesos, é comum realizar a normalização dos vetores para evitar que documentos mais longos tenham maior influência nos resultados. Essa normalização, geralmente, é feita utilizando-se a norma do vetor:

$$\vec{d}_{norm} = \frac{\vec{d}}{||\vec{d}||} \tag{2.4}$$

Normalmente, emprega-se o cálculo do cosseno para determinar o nível de semelhança entre a consulta e o documento (SOUZA *et al.*, 2017). Assim, a similaridade é proporcional ao cosseno do ângulo entre o vetor que representa o documento e o vetor da consulta, sendo calculada por:

$$\text{sim}(d,q) = \frac{\vec{d} \cdot \vec{q}}{||\vec{d}||\,||\vec{q}||} \tag{2.5}$$

onde $\vec{d} \cdot \vec{q}$ representa o produto escalar entre os vetores, e $||\vec{d}||$ e $||\vec{q}||$ corres-

pondem às suas magnitudes. Quanto mais próximo de 1 for o valor da similaridade, maior será a relevância do documento em relação à consulta.

2.1.1.2 Modelo Booleano

O modelo booleano é uma das técnicas mais tradicionais e simples utilizadas na RI. Ele foi proposto nos anos 1960 e utiliza lógica booleana para representar documentos e consultas (Souza, 2006). A ideia central é que o usuário constrói consultas usando operadores lógicos como *AND*, *OR* e *NOT*, combinando palavras-chave de forma precisa (Ferneda, 2003). A recuperação, nesse caso, funciona de maneira binária, ou seja, ou o documento satisfaz a consulta, ou não satisfaz.

O processo funciona da seguinte forma: quando um usuário deseja encontrar informações em um sistema baseado no modelo, ele formula uma expressão lógica contendo termos relacionados à busca. Por exemplo, a consulta "computação AND inteligência OR dados AND aprendizado"retornará documentos que contenham, obrigatoriamente, as palavras “computação” e “inteligência” ou “dados” e “aprendizado”. Essa estrutura fornece controle direto sobre o que será recuperado, mas exige que o usuário saiba como montar as expressões corretamente.

Apesar de sua simplicidade, o modelo booleano apresenta algumas limitações importantes. Ele não oferece um sistema de ranqueamento para indicar quais documentos são mais relevantes em relação à consulta; todos os resultados são tratados de maneira igual (Souza, 2006). Além disso, consultas mal formuladas podem excluir documentos importantes ou trazer resultados excessivamente amplos. Não há tratamento para sinônimos ou contexto semântico, o que dificulta buscas mais refinadas em sistemas maiores.

Mesmo com essas limitações, o modelo booleano foi essencial para o desenvolvimento inicial da RI e ainda é utilizado em alguns sistemas simples ou como base para outros modelos mais avançados.

2.1.1.3 Modelo Fuzzy

O modelo fuzzy foi desenvolvido como uma alternativa aos modelos tradicionais, como o booleano, buscando lidar melhor com a imprecisão e incerteza inerentes à linguagem humana (Ferneda; Dias, 2013). Esse modelo utiliza conceitos da lógica fuzzy (ou lógica difusa), proposta por Lotfi Zadeh em 1965, que permite trabalhar com graus de pertencimento em vez de decisões binárias (Souza, 2006). Enquanto o modelo booleano classifica um documento de forma binária, o modelo fuzzy permite atribuir um grau de relevância contínuo, geralmente entre 0 e 1, a cada documento recuperado (Souza, 2006).

No modelo fuzzy, documentos e consultas são representados por conjuntos fuzzy de termos (Ferneda, 2003). A similaridade entre um documento e uma consulta não é apenas uma correspondência exata de palavras, mas é calculada com base em um grau de pertinência (Souza, 2006). Assim, um documento pode ser recuperado mesmo que não contenha exatamente todos os termos especificados, desde que possua um conteúdo considerado suficientemente relacionado com a consulta (Souza, 2006). Isso permite uma maior flexibilidade na busca e resultados mais próximos da intenção do usuário.

Uma das principais vantagens do modelo fuzzy é a capacidade de lidar com imprecisão e ambiguidade na formulação de consultas (Ferneda, 2003). Em vez de exigir que o usuário saiba exatamente quais termos usar, o sistema pode retornar documentos que estejam relacionados de forma parcial ou aproximada. Além disso, os resultados podem ser classificados por grau de relevância, ajudando o usuário a identificar os documentos mais importantes para sua busca.

Por outro lado, a implementação do modelo fuzzy tende a ser mais complexa do que a de modelos mais simples, como o booleano (Ferneda; Dias, 2013). A definição dos graus de pertinência, a escolha de funções de similaridade e a configuração de parâmetros fuzzy exigem um cuidado maior no desenvolvimento do sistema. Mesmo assim, essa abordagem é muito útil em contextos em que a busca precisa ser mais tolerante a variações linguísticas e subjetividade, como em bancos de dados complexos e sistemas de RI na *web* (Ferneda; Dias, 2013).

2.1.2 Modelos Dinâmicos

Os modelos quantitativos de RI representam os documentos por meio de termos de indexação e pesos, de forma impositiva e sem participação do usuário. Em contrapartida, os modelos apresentados nesta seção destacam a importância da intervenção do usuário nesse processo. Por meio da interação e do *relevance feedback*, os usuários ajudam a adaptar e refinar a representação dos documentos de acordo com seus interesses e necessidades, melhorando a qualidade da recuperação das informações (Ferneda, 2003).

2.1.2.1 Sistemas especialistas

Os sistemas especialistas podem ser aplicados na RI para fornecer respostas mais qualificadas em áreas específicas, atuando como filtros inteligentes entre a consulta do usuário e o banco de dados. Em sistemas de RI, especialmente nos casos que exigem conhecimento especializado, o sistema especialista auxilia na interpretação de consultas complexas e na seleção das respostas mais apropriadas (Giarratano

et al., 2005). Essa abordagem é eficaz em contextos restritos e bem definidos, como sistemas médicos ou jurídicos, onde é necessário associar termos técnicos a documentos específicos (RUSSELL *et al.*, 2009).

2.1.2.2 Redes neurais

As redes neurais artificiais têm papel crescente na RI, sobretudo, com o avanço das técnicas de aprendizado profundo. Elas são utilizadas para analisar grandes volumes de dados e melhorar a correspondência entre consultas e documentos, captando relações semânticas além da simples correspondência de palavras-chave (Goodfellow *et al.*, 2016). As redes neurais são essenciais em sistemas modernos de busca e recomendação, permitindo o desenvolvimento de motores de busca mais inteligentes e personalizados (Haykin, 2009). Além disso, redes neurais profundas têm sido aplicadas em tarefas de classificação de documentos e filtragem colaborativa em RI.

2.1.2.3 Algoritmos genéticos

Os *algoritmos genéticos* também encontram aplicações na RI, especialmente no contexto de otimização de consultas e ajustamento dos pesos de indexação dos termos. Esses algoritmos podem ser utilizados para refinar as consultas do usuário, buscando melhores combinações de palavras-chave ou termos que maximizem a relevância dos documentos recuperados (Goldberg, 1989). Além disso, algoritmos genéticos são úteis na personalização de sistemas de RI e na implementação de mecanismos de *relevance feedback*, adaptando os resultados recuperados ao perfil e aos interesses do usuário (Eiben *et al.*, 01/2003).

2.2 Estado da arte

(Ferneda, 2003) analisa a contribuição da Computação para a RI na CI. O trabalho revisa criticamente modelos computacionais e técnicas de Processamento de Linguagem Natural (PLN), ressaltando limitações conceituais e a necessidade de integração mais profunda entre as áreas.

(Mitra *et al.*, 2018) discutem os avanços em RI com modelos neurais, organizando-os em *embeddings* distribuídos, modelos supervisionados e arquiteturas profundas. Destacam desafios como custo computacional e exigência de dados rotulados, sugerindo abordagens híbridas como futuro promissor.

(Devlin *et al.*, 2019) apresentam o *Bidirectional Encoder Representations from Transformers* (BERT), modelo pré-treinado baseado em *Transformers*, com pré-

treinamento por *Masked LM* (MLM) e *Next Sentence Prediction* (NSP). O BERT supera métodos anteriores em várias tarefas de PLN, e os autores indicam continuidade das pesquisas em arquiteturas mais profundas e estratégias diversificadas.

(Vivekavardhan, 12/2015) analisa o papel dos mecanismos de busca na RI, com foco na comunicação científica. A introdução destaca a evolução desses sistemas e a necessidade de um design centrado no usuário diante da crescente complexidade informacional. Metodologicamente, a pesquisa realiza uma revisão teórica sobre práticas de RI, abordando aspectos técnicos, comportamentais e sociais. O estudo ressalta limitações como a cobertura parcial da *web* e a necessidade de múltiplas fontes, concluindo pela importância de soluções personalizadas e integradas, apontando tendências como pesquisa por voz e uso de tecnologias emergentes.

Por fim, este trabalho configura-se como um estudo histórico sobre os modelos de busca, além de se apresentar como um referencial para a realização de pesquisas futuras.

3 METODOLOGIA

Este capítulo abordará a classificação do trabalho na Seção 3.1, seguida da apresentação dos princípios definidos pelo protocolo de revisão sistemática da literatura, cuja aplicação será demonstrada na Seção 3.2. Por fim, a Seção 3.3 trará um detalhamento sobre a adaptação da metodologia ágil Scrum no contexto deste estudo.

3.1 Classificação da pesquisa

Segundo (Moresi, 2003), esta pesquisa é classificada conforme os seguintes critérios. Quanto à sua natureza, trata-se de uma pesquisa básica, pois tem como objetivo aprofundar o conhecimento teórico sobre motores de busca na RI, sem aplicação prática imediata dos resultados. No que se refere à forma de abordagem, a pesquisa é caracterizada como qualitativa, uma vez que se baseia na análise e interpretação de conceitos, modelos e abordagens teóricas presentes na literatura, sem a realização de tratamento estatístico dos dados.

Em relação aos fins, configura-se como uma pesquisa descritiva, uma vez que visa identificar e caracterizar métodos, tecnologias e critérios empregados nos motores de busca ao longo do tempo, sem necessariamente explicar as causas dos fenômenos observados. Quanto aos meios de investigação, é classificada como uma pesquisa bibliográfica e documental, pois foi efetuada uma revisão sistemática baseada em fontes acessíveis ao público, como artigos científicos, livros e bases de dados acadêmicas.

De acordo com (Wazlawick, 2009), este trabalho se enquadra no estilo de pesquisa voltado à apresentação e comparação de abordagens distintas. No caso específico, o estudo comparou estratégias, métricas e evoluções metodológicas relacionadas aos motores de busca utilizados na RI, contribuindo, assim, para uma visão crítica e fundamentada sobre a área.

3.2 Protocolo de revisão sistemática da literatura

Com o objetivo de orientar o desenvolvimento deste trabalho e servir como base para a construção do método de pesquisa adotado nesta Revisão Sistemática da Literatura (RSL), definiu-se a aplicação de um protocolo de RSL. O protocolo selecionado foi proposto pelos autores (Okoli *et al.*, 2010), e estabelece uma estrutura composta por oito tópicos que devem ser seguidos para garantir que a pesquisa possua uma base teórica consistente, um método confiável e resultados satisfatórios ao final. Este protocolo é voltado especificamente para a área de Sistemas de Informação

(SI).

De acordo com (Okoli *et al.*, 2010), o primeiro passo na aplicação do protocolo consiste na definição do problema que será investigado na pesquisa. Neste trabalho, a pergunta central, que orientou o desenvolvimento foi: “Historicamente, quais os principais motores de busca e quais seus conceitos?”. Como pergunta secundária, tem-se: “Quais artigos podem ser citados como exemplos de aplicações destes motores e quais seus objetivos?”. Considerando o espaço amostral previamente estabelecido, as técnicas foram organizadas conforme sua data de surgimento, com a apresentação de seus conceitos por meio de exemplos e imagens que facilitassem a compreensão. Também foram destacadas suas aplicações mais recentes, com base em fontes como artigos, teses, monografias e livros. Os veículos de pesquisa utilizados incluíram Scielo, Google Acadêmico, Capes, repositórios de instituições como Universidade de São Paulo (USP), Instituto Federal de São Paulo (IFSP), Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal de Santa Catarina (UFSC), Instituto Federal de Ciência e Tecnologia de Minas Gerais (IFMG), além de livros relevantes, artigos e revistas científicas.

Como segundo ponto, os autores (Okoli *et al.*, 2010) destacam a importância de apresentar com total clareza, precisão e riqueza de detalhes o procedimento a ser adotado. Para garantir que esta RSL atenda aos critérios de credibilidade exigidos, adotou-se um protocolo específico de RSL, no qual foram descritos tanto os métodos utilizados quanto os resultados obtidos ao longo da investigação.

O terceiro tópico deste protocolo trata da estratégia de busca pela literatura que fundamentou este trabalho. Para assegurar que as fontes selecionadas fossem confiáveis e estivessem alinhadas com os critérios estabelecidos, explicitaram-se os métodos de pesquisa adotados, mantendo-se a fidelidade tanto à abrangência do tema quanto ao espaço amostral previamente definido. A fim de obter conhecimentos relevantes para a compreensão da proposta, foram utilizadas as seguintes palavras-chave: Motores de Busca, Recuperação de Informação, Algoritmos de Busca, Inteligência Artificial em Motores de Busca e outros. Para a busca dos artigos abordados, pesquisou-se o termo “Motores de Busca” na plataforma Google Scholar. Para os artigos que serviram como base para exemplificação, efetuou-se a busca pelo termo “Modelos de Busca” na mesma plataforma.

A quarta etapa desta pesquisa corresponde à triagem prática do conteúdo utilizado. Nessa fase, são explicitados, ao longo da escrita, os métodos de trabalho e os materiais de estudo adotados, bem como os critérios de exclusão e seleção das fontes. Foram considerados os motores de busca e modelos teóricos que apresentam maior recorrência na literatura, identificados por meio de bases acadêmicas consolidadas, como o Google Scholar, levando em conta o número de citações dos trabalhos relacionados, suas contribuições inovadoras em relação a tecnologias anteriores e sua

relevância no desenvolvimento da área. A análise foi baseada em materiais consolidados, publicados no intervalo entre 1990 e 2020. Foram excluídos artigos fora desse recorte temporal ou que, após a leitura de títulos e resumos, não se mostraram alinhados aos objetivos da pesquisa. Da mesma forma, foram desconsiderados motores de busca e abordagens com baixa recorrência em estudos acadêmicos ou com reduzido número de citações quando comparados a outras soluções equivalentes.

O quinto passo consiste na aplicação de uma avaliação de qualidade, com o objetivo de determinar quais artigos possuem relevância suficiente para serem incluídos na síntese da RSL. Para isso, foi elaborado e aplicado um teste de relevância voltado à análise dos estudos selecionados. Os artigos foram pontuados com base na qualidade das metodologias de pesquisa utilizadas e nos resultados apresentados, conforme os critérios definidos em tabela específica. Como etapa inicial de triagem, considerou-se a quantidade de citações de cada artigo encontrado. Em seguida, foi efetuada a leitura dos títulos, resumos e *abstracts*, a fim de selecionar os estudos mais relevantes, de acordo com a pontuação obtida, adotando-se uma abordagem qualitativa para a seleção final.

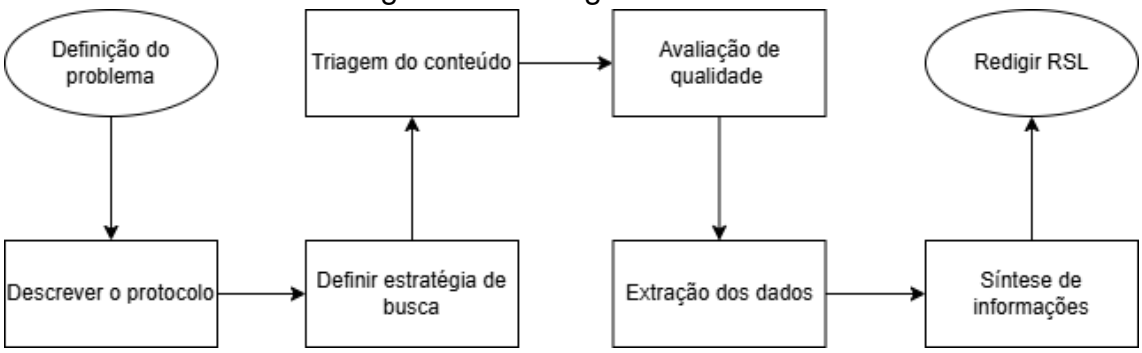
Após a realização da avaliação de qualidade, foi conduzida a extração dos dados relevantes à pesquisa como sexto passo. Com os estudos selecionados e seus pontos de interesse devidamente identificados, extraíram-se as informações de forma sistemática, considerando: a conceituação de cada motor de busca analisado, exemplos de aplicação em diferentes contextos e métricas de desempenho utilizadas em suas avaliações.

Como sétimo ponto deste protocolo, efetuou-se a síntese das informações coletadas nas etapas anteriores. Nessa fase, os dados extraídos foram analisados e integrados, utilizando-se as técnicas previamente estabelecidas, de modo a apresentar os resultados de forma clara, objetiva e coerente com os objetivos do trabalho.

O oitavo e último passo consiste na redação da RSL. Com base em todo o processo desenvolvido, elaborou-se a monografia, detalhando as informações obtidas e destacando os aspectos mais relevantes, os critérios de importância definidos ao longo da pesquisa. O texto foi estruturado com o objetivo de servir como material de referência e estudo na área de motores de busca, contribuindo para o avanço do conhecimento sobre o tema.

De acordo com a Figura 1, observa-se uma visão geral em alto nível que auxilia na compreensão integrada de todas as etapas do protocolo de RSL.

Figura 1 – Fluxograma da RSL



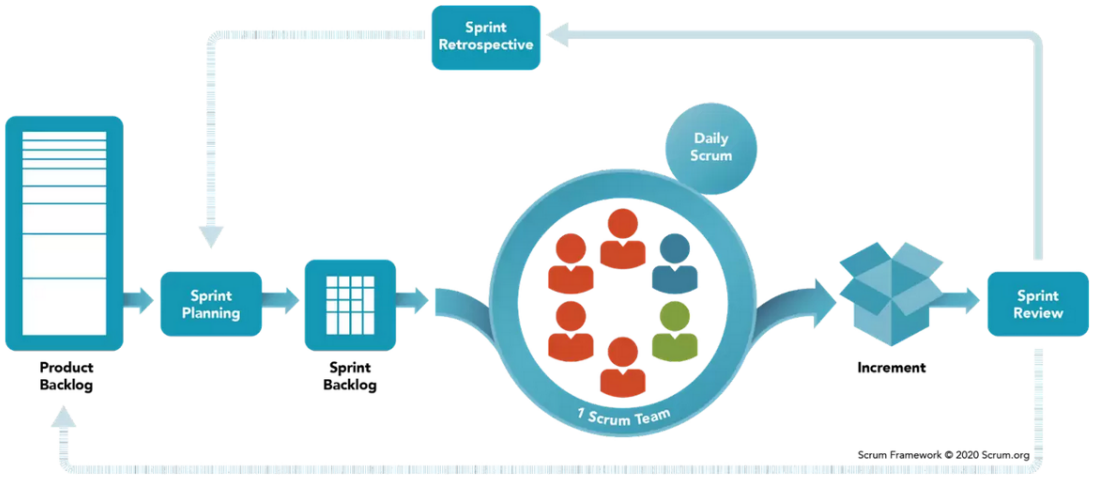
Fonte: Elaborado pelo próprio autor, 2025.

3.3 Scrum

Scrum é um *framework* ágil de gerenciamento de projetos, com foco na colaboração, adaptação e entrega contínua de valor. De acordo com (Schwaber *et al.*, 2002), criador do *scrum*, trata-se de uma estrutura que não determina ações específicas para cada situação, mas fornece um conjunto de práticas que facilitam a visualização e o acompanhamento do projeto.

Neste projeto, efetuou-se uma adaptação da metodologia ágil *Scrum*, Figura 2, contando com a participação do estudante orientado e do docente orientador como membros da equipe. De acordo com (Schwaber *et al.*, 2002), um dos elementos essenciais são os encontros nos quais se avalia o que já foi feito e se definem os próximos passos a serem seguidos. Para este trabalho, esses encontros ocorreram semanalmente, com a presença da equipe mencionada. Um dos artefatos dessa metodologia é o *product backlog*. O *product backlog* 1 corresponde ao conjunto de entregas obtidas a partir das tarefas previamente estabelecidas para execução ao longo do projeto. Outro artefato importante são as *sprints* 2, nas quais as tarefas serão organizadas e distribuídas conforme sua relevância, visando otimizar o andamento do trabalho. Essa organização foi feita de acordo com a disponibilidade de tempo para cada atividade programada para a equipe. As *sprints* estão elencadas das Tabelas 1 a 6.

Figura 2 – Funcionamento Scrum



Fonte: Scrum.org (2020)

Tabela 1 – Backlog do projeto

ID	Atividade	Importância	Estimativa
1	Embasamento teórico	10	8 meses
2	Definição do problema	10	1 mês
3	Montar teste de relevância	10	1 mês
4	Definir base de dados, palavras-chave e espaço amostral	10	1 mês
5	Aplicar teste de relevância	10	2 meses
6	Discussão dos resultados	10	2 meses
7	Escrita da monografia	10	8 meses
8	Defesa da monografia	10	1 mês

Fonte: Elaborado pelo autor, 2025.

Tabela 2 – *Sprint* 1

ID	Atividade	Importância	Estimativa
1	Embasamento teórico	10	8 meses

Fonte: Elaborado pelo autor, 2025.

Tabela 3 – *Sprint* 2

ID	Atividade	Importância	Estimativa
2	Definição do problema	10	1 mês
3	Montar teste de relevância	10	1 mês
4	Definir base de dados, palavras-chave e espaço amostral	10	1 mês

Fonte: Elaborado pelo autor, 2025.

Tabela 4 – *Sprint* 2

ID	Atividade	Importância	Estimativa
5	Aplicar teste de relevância	10	2 meses
6	Discussão dos resultados	10	2 meses

Fonte: Elaborado pelo autor, 2025.

Tabela 5 – *Sprint* 3

ID	Atividade	Importância	Estimativa
7	Escrita da monografia	10	8 meses

Fonte: Elaborado pelo autor, 2025.

Tabela 6 – *Sprint* 3

ID	Atividade	Importância	Estimativa
8	Defesa da monografia	10	1 mês

Fonte: Elaborado pelo autor, 2025.

3.4 Materiais e tecnologias

Para a realização da pesquisa e escrita deste trabalho, empregou-se um computador do aluno orientado, sendo um notebook com um processador Intel Core i5-10210, oito *gigabytes* de memória RAM.

4 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos após a realização da pesquisa, respeitando os princípios propostos pelo protocolo de RSL utilizado, juntamente com exemplos e figuras que possam auxiliar a elencar fatores importantes e facilitar a compreensão de seus leitores. Para o desenvolvimento deste trabalho, foram adotados como ponto de estudo 7 motores de busca que apresentaram mudanças e/ou resultados significativos em relação a modelos anteriores. Para a realização da RSL, foi conduzido um teste de relevância em aproximadamente 142 artigos previamente selecionados, dos quais 78 foram considerados adequados e utilizados no desenvolvimento da pesquisa.

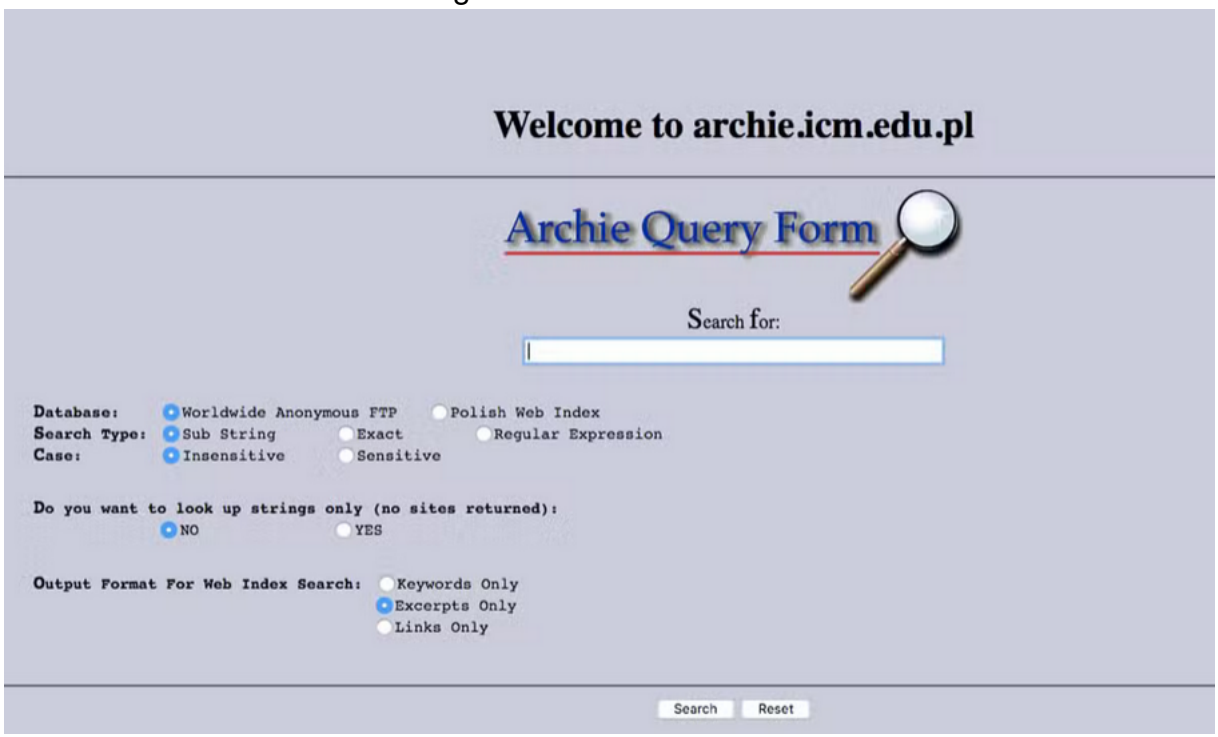
4.1 Análise do Modelo de Funcionamento do Motor Archie

O *Archie*, Figura 3, desenvolvido em 1990 por Alan Emtage, então estudante da Universidade McGill, no Canadá, é reconhecido como o primeiro sistema automatizado de indexação e busca na internet, sendo considerado o antecessor direto dos motores de busca modernos (Tavares *et al.*, 2009). À época, a *World Wide Web* (WWW) ainda não estava amplamente difundida, e o compartilhamento de dados era realizado majoritariamente por meio de servidores FTP públicos (Cendón, 2001). Nesse ambiente, a localização de arquivos dependia do conhecimento prévio da estrutura e do Localizador Uniforme de Recurso (URL) dos servidores, o que tornava o processo manual, ineficiente e suscetível a erros.

O *Archie* foi concebido para mitigar esse problema citado no parágrafo anterior por meio da automação da coleta e indexação de nomes de arquivos hospedados nesses servidores (Seymour *et al.*, 2011). Utilizando agentes automatizados, o sistema realizava varreduras periódicas nos diretórios públicos de servidores FTP, extraindo os metadados e armazenando-os em uma base de dados centralizada (Escandell-Poveda *et al.*, 2022). As consultas eram realizadas por meio de interface textual, com suporte a Telnet ou comandos em linha, retornando os endereços dos servidores que continham os arquivos correspondentes aos termos buscados.

Apesar de, em sua implementação original, não realizar a indexação do conteúdo interno dos arquivos nem dispor de uma interface gráfica, o *Archie* introduziu conceitos fundamentais da RI em ambientes distribuídos: varredura automatizada, indexação periódica, busca por palavras-chave e resposta programada ao usuário (Seymour *et al.*, 2011). Com o tempo, versões posteriores passaram a incorporar representações visuais ou interfaces *web* básicas para facilitar o acesso ao sistema, como ilustrado na Figura 3. O modelo conceitual do *Archie* serviu de base para sistemas subsequentes, como Veronica e Jughead, e estabeleceu os princípios que viriam

Figura 3 – Sistema Archie



Fonte: Stackscale (2021).

a ser amplamente explorados por motores de busca como Yahoo e Google (Escandell-Poveda *et al.*, 2022).

4.1.1 Aplicações Práticas do Archie

Mesmo com suas limitações tecnológicas, o *Archie* desempenhou um papel importante no cotidiano de pesquisadores, estudantes e profissionais de tecnologia durante os anos 1990. Suas principais aplicações práticas incluíam:

- **Distribuição de software livre:** Muitos desenvolvedores utilizavam o *Archie* para localizar versões atualizadas de programas de código aberto, como o sistema operacional Linux, compiladores GNU Compiler Collection (GCC), bibliotecas Unix, entre outros (Ferneda, 2003).
- **Localização de drivers e utilitários:** Usuários podiam encontrar arquivos de configuração, drivers para dispositivos e ferramentas de diagnóstico hospedadas em servidores de universidades e organizações técnicas (Escandell-Poveda *et al.*, 2022).
- **Busca por documentação técnica:** Diversas instituições mantinham manuais, tutoriais, dissertações e artigos disponíveis via FTP. O *Archie* permitia que esses documentos fossem encontrados de maneira mais ágil (Ferneda, 2003).

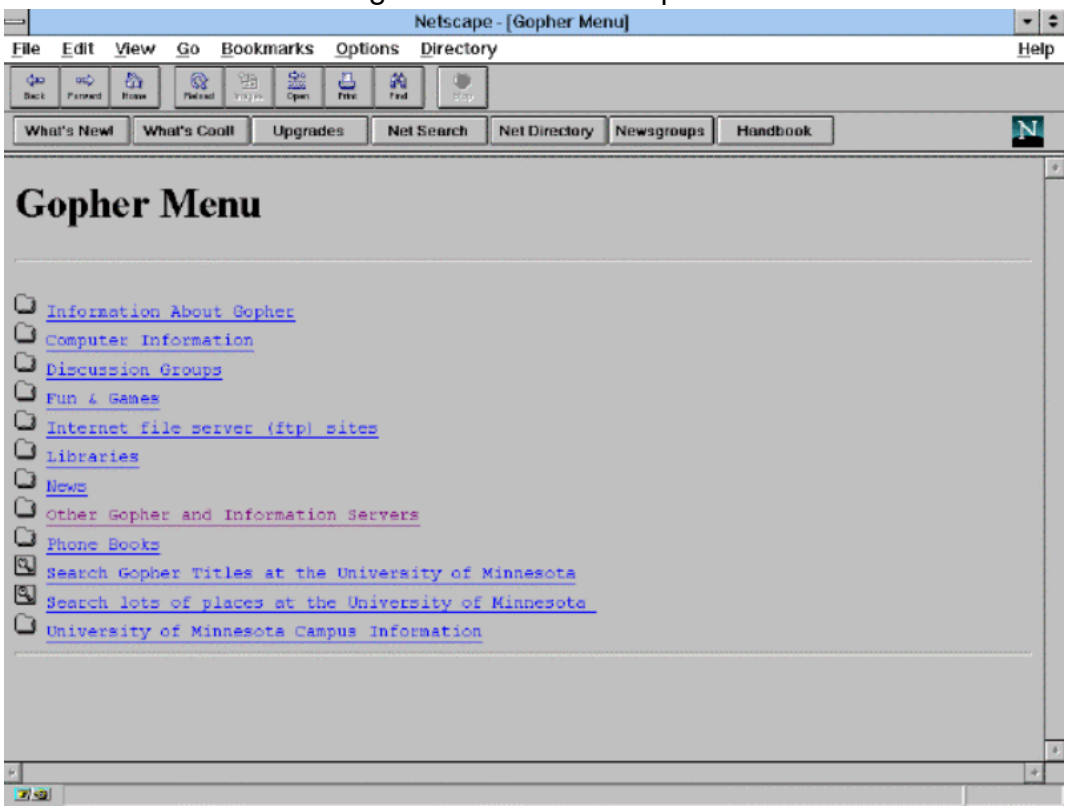
- **Acesso a atualizações de sistemas:** Patches e arquivos de atualização de sistemas operacionais também eram distribuídos por FTP e podiam ser localizados rapidamente usando o *Archie* (Ferneda, 2003).

Essas aplicações demonstram como o *Archie* antecipou a importância dos mecanismos de busca para a organização e o acesso à informação digital. Ele não apenas facilitou a vida dos usuários da internet naquela época, como também consolidou um modelo de interação com grandes volumes de dados distribuídos, abrindo caminho para a evolução dos sistemas de busca contemporâneos.

4.2 Análise do Modelo de Busca Gopher

O *Gopher* 4, desenvolvido em 1991 por Mark McCahill e sua equipe, na Universidade de Minnesota, foi um dos primeiros sistemas projetados para organizar e disponibilizar informações na *internet* (PICALHO, 2023). Diferentemente do *Archie*, que se limitava a indexar listagens de arquivos armazenados em servidores FTP (Seymour *et al.*, 2011), o *Gopher* introduziu uma estrutura hierárquica de diretórios e menus (Anklesaria *et al.*, 1993). Essa organização permitia ao usuário navegar intuitivamente entre conteúdos distribuídos em servidores distintos, representando um avanço significativo na forma de acessar documentos, artigos e materiais acadêmicos (Ferneda, 2003).

Figura 4 – Sistema Gopher



Fonte: Showmetech, (2019).

Do ponto de vista técnico, o protocolo *Gopher* utilizava comunicação via *Transmission Control Protocol / Internet Protocol* (TCP/IP), operando, por padrão, na porta 70 (Jara Morante, 1994). O cliente enviava ao servidor uma *selector string* o (caminho do recurso solicitado) e recebia de volta um menu ou documento em texto simples (Anklesaria *et al.*, 1993). Esse funcionamento foi descrito formalmente na *Request for Comments* (RFC), 1436, publicada em 1993 pela *Internet Engineering Task Force* (IETF), responsável por padronizar as especificações técnicas do protocolo (Anklesaria *et al.*, 1993). Cada item de menu era associado a um tipo de recurso específico, conforme exemplificado na Tabela 7, acompanhado de informações como endereço do servidor e porta de acesso.

Tabela 7 – Itens do Menu

ID	Descrição
0	Arquivos de texto
1	Submenus
7	Buscas

Fonte: Elaborado pelo autor, 2025.

A principal inovação do *Gopher* foi a criação do *Gopherspace*, um modelo de navegação estruturado que rapidamente se expandiu em universidades, bibliotecas digitais e centros de pesquisa. Inicialmente, porém, o sistema não possuía mecanismos de busca integrados, o que levou ao desenvolvimento de ferramentas complementares, como o Veronica (1992), que indexava globalmente títulos de menus, e o *Jughead* (1993), especializado em buscas locais (PICALHO, 2023). Assim, o *Gopher* não apenas organizou o acesso à informação, como também pavimentou o caminho para o surgimento de motores de busca mais sofisticados.

A partir de 1993, entretanto, sua popularidade declinou rapidamente devido ao crescimento da *World Wide Web* (WWW), que oferecia maior flexibilidade por meio do HTML, suporte a elementos gráficos e interfaces mais amigáveis. A decisão da Universidade de Minnesota, de adotar um modelo de licenciamento para o *Gopher*, também reduziu sua adoção (Anklesaria *et al.*, 1993), enquanto a WWW foi disponibilizada em domínio público pelo CERN, favorecendo sua disseminação global. Com isso, o *Gopher* tornou-se, gradualmente, uma tecnologia de nicho, hoje reconhecida mais por seu papel histórico do que por uso ativo.

Ainda assim, seu legado permanece significativo. O *Gopher* demonstrou o potencial de sistemas distribuídos de organização de informações e antecipou conceitos posteriormente consolidados nos navegadores *web* e motores de busca contemporâneos. Sua arquitetura simples continua sendo referência em estudos sobre protocolos de rede e evolução da comunicação digital.

4.2.1 Aplicações Práticas do Gopher

Algumas das aplicações práticas mais importantes do Gopher ocorreram em bibliotecas acadêmicas e sistemas de informação nas universidades. Por exemplo, bibliotecários da Indiana University criaram servidores Gopher para disponibilizar catálogos e outras informações da comunidade universitária de forma hierárquica e navegável (Kosokoff, 1995).

Na *University College London* (UCL), o *Gopher* foi usado para prover acesso a múltiplos serviços: notícias do centro de computação, catálogo da biblioteca, diretório telefônico e até acesso Telnet ao sistema da biblioteca (SHINTAKU *et al.*, 1995).

Outro caso prático envolve a biblioteca de ciências da vida da Penn State, que organizou seu servidor Gopher por temas *(“Agriculture and Biology”, “Health and Biomedical”)* e usava esse canal para divulgar informação efêmera, como chamadas para bolsas, vagas e eventos acadêmicos (McMillan *et al.*, 1997).

Também há relatos de uso institucional em faculdades menores, como no Middlebury College, onde bibliotecários e pessoal de Tecnlogia da Informação (TI) co-

laboraram para instalar um servidor Gopher como parte de um sistema de cooperação de informação (AL-KHULAIFI, 1997).

No contexto mais amplo, o Gopher era como uma “cola eletrônica” que integrava catálogos de bibliotecas, bases de dados via WAIS, FTP e Telnet, permitindo que diferentes tipos de recursos de informação fossem acessados de forma unificada (CIOLEK, 1998).

4.2.2 Vantagens e Desvantagens

O *Gopher* representou um marco nos primeiros esforços de organização sistemática da informação na *internet*, oferecendo uma navegação simples e hierárquica que superava soluções anteriores, como o *Archie* (Anklesaria *et al.*, 1993). Apesar de seu impacto inicial, especialmente em ambientes acadêmicos, o sistema apresentava limitações que se tornariam mais evidentes com o surgimento da WWW (BAEZA-YATES *et al.*, 1999).

4.2.2.1 Vantagens

- Estrutura hierárquica simples e intuitiva, facilitando a navegação em comparação com listas de arquivos, como no *Archie* (Anklesaria *et al.*, 1993);
- Padronização definida pela RFC 1436, permitindo interoperabilidade entre clientes e servidores (Anklesaria *et al.*, 1993);
- Baixo consumo de banda, adequado à infraestrutura limitada da internet no início dos anos 1990 (Anklesaria *et al.*, 1993);
- Forte adoção em universidades e centros de pesquisa, possibilitando repositórios acessíveis globalmente (SHINTAKU *et al.*, 1995).

4.2.2.2 Desvantagens

- Dependência de menus hierárquicos fixos, o que dificultava a localização de informações em grandes volumes de dados (SHINTAKU *et al.*, 1995);
- Ausência inicial de mecanismos de busca integrados (BAEZA-YATES *et al.*, 1999);
- Baixa flexibilidade para incorporar diferentes formatos de mídia, como imagens e hiperlinks livres (BERNERS-LEE *et al.*, 1994);
- Perda de competitividade com o surgimento da WWW, mais rica, flexível e amplamente distribuída (CIOLEK, 1998).

4.3 O Surgimento do Veronica no Gopherspace

O Veronica foi desenvolvido em 1992 por Steve Foster e Fred Barrie, na Universidade de Nevada, com o objetivo de suprir uma limitação essencial do protocolo *Gopher*: a ausência de um mecanismo de busca global (VERONICA:. . . , 1990). Enquanto o *Gopher* permitia apenas a navegação por menus e diretórios (Jara Morante, 1994), o Veronica surgiu como um índice capaz de pesquisar títulos de menus em todos os servidores *Gopher* disponíveis na rede (VERONICA:. . . , 1990). Dessa forma, transformou o *Gopherspace* em um ambiente muito mais acessível, aproximando-se daquilo que hoje entendemos por motores de busca.

Do ponto de vista técnico, o Veronica funcionava como uma base de dados centralizada, atualizada periodicamente com informações coletadas dos servidores *Gopher* (BOS, 1997). Quando um usuário submetia uma consulta por palavra-chave, o sistema retornava uma lista de menus e diretórios relevantes, permitindo acesso direto aos conteúdos correspondentes (VERONICA:. . . , 1990). A integração entre o protocolo *Gopher* e o Veronica garantiu uma experiência de navegação mais dinâmica, reduzindo a necessidade de explorar manualmente estruturas hierárquicas extensas (WIGGINS, 1993).

A principal inovação do Veronica foi introduzir o conceito de busca textual em escala global, antecipando práticas que mais tarde seriam consolidadas nos motores de busca da *web* (WIGGINS, 1993). Embora restrito ao universo do *Gopherspace*, sua relevância foi notável, pois demonstrou a importância da indexação automatizada como estratégia de organização da informação (BOS, 1997). Pouco tempo depois, surgiram variações como o *Jughead*, voltado para buscas locais em servidores específicos, evidenciando que o futuro da navegação digital dependia cada vez mais de sistemas capazes de pesquisar e organizar conteúdos de forma eficiente.

4.3.1 O Processo de Indexação Automatizada

A indexação realizada pelo Veronica pode ser classificada como automatizada, pois não exigia intervenção humana para catalogar individualmente cada item. Esse processo era conduzido por um programa (um tipo de rastreador primitivo) que percorria sistematicamente o *Gopherspace* de forma autônoma (IETF RFC Editor, 1994).

O rastreador do *Veronica* tinha como função percorrer sistematicamente os servidores *Gopher* registrados, coletando e indexando dois tipos principais de informações de cada item disponível nos menus (WIGGINS, 1993).

a) **O título do item:** A descrição textual visível para o usuário no menu (WIGGINS, 1993);

b) **A selector string e o host:** O endereço técnico e a cadeia de seleção que permitiam ao sistema localizar e recuperar o conteúdo correspondente (WIGGINS, 1993).

Essa abordagem discordava fundamentalmente da indexação manual, comum em bibliotecas digitais da época, onde metadados eram atribuídos por especialistas (VERONICA:. . . , 1990). O caráter automatizado e abrangente do Veronica permitiu a construção de um índice unificado de todo o conteúdo disponível, algo inédito para a época (VERONICA:. . . , 1990).

4.3.2 Limitações da Indexação e Atualização

Embora fosse automatizada, a indexação do Veronica apresentava limitações notáveis em comparação com os motores de busca atuais:

- **Foco Apenas em Títulos:** A busca e a indexação eram realizadas exclusivamente sobre o texto presente no título dos itens de menu. O sistema não lia o conteúdo integral dos documentos (como arquivos de texto ou binários) a que os menus apontavam. Isso significava que um documento cujo conteúdo era relevante, mas cujo título era genérico, não seria facilmente encontrado (WIGGINS, 1993);
- **Atualização Periódica:** O índice do Veronica não era atualizado em tempo real. Os rastreamentos eram realizados em intervalos de tempo predeterminados (periodicamente), o que significava que servidores *Gopher* recém-criados ou alterações recentes nos menus demoravam a ser incorporadas ao banco de dados de busca (WIGGINS, 1993);
- **Restrição ao Protocolo Gopher:** A indexação estava restrita a servidores que utilizavam o protocolo *Gopher*. Com o advento e a popularidade crescente da *web* e do protocolo *Hypertext Transfer Protocol* (HTTP), o Veronica tornou-se rapidamente obsoleto, uma vez que não era projetado para rastrear e indexar o novo formato de páginas HTML (BOS, 1997).

4.3.3 Vantagens e Desvantagens

O Veronica surgiu como uma resposta direta à principal limitação do *Gopher*: a ausência de mecanismos de busca eficientes (SHINTAKU *et al.*, 1995). Ao oferecer um sistema de indexação global de menus, tornou a navegação mais dinâmica e aproximou a experiência dos usuários daquilo que hoje conhecemos como motores de busca (WIGGINS, 1993). Apesar de sua inovação e popularidade inicial, o Veronica também encontrou barreiras técnicas e contextuais que restringiram sua

continuidade, especialmente diante do avanço acelerado da *web* e de seus motores de busca próprios (BERNERS-LEE *et al.*, 1994; CIOLEK, 1998).

4.3.3.1 Vantagens do Veronica na Época

As principais vantagens do Veronica na época foram:

- **Busca em Larga Escala:** Foi o primeiro sistema de busca em larga escala dentro do *Gopherspace*, superando a limitação da navegação manual e hierárquica do *Gopher* (WIGGINS, 1993);
- **Recuperação por Palavras-Chave:** Permitia consultas por palavras-chave, tornando a RI muito mais rápida e eficiente para o usuário (WIGGINS, 1993);
- **Maior Dinamismo:** Ampliou a utilidade do *Gopher*, tornando-o mais dinâmico e próximo da experiência de busca que hoje associamos a motores como Google (WIGGINS, 1993).

4.3.3.2 Desvantagens do Veronica na Época

As principais desvantagens do Veronica na época foram:

- **Limitação Protocolar:** Limitado exclusivamente ao universo do *Gopher*, não abrangia outros protocolos em crescimento, como FTP ou, posteriormente, HTTP (WIGGINS, 1993);
- **Indexação Superficial:** Indexava apenas títulos de menus e diretórios, sem acesso ao conteúdo interno dos documentos, o que resultava em busca de qualidade variável (WIGGINS, 1993);
- **Gargalos de Disponibilidade:** Dependia de servidores centrais de indexação, o que frequentemente gerava gargalos de atualização e problemas de disponibilidade para os usuários (WIGGINS, 1993).

4.4 Jughead

O *Jughead* foi desenvolvido em 1993 por Rhett Jones, na Universidade de Utah, como uma solução complementar ao sistema *Gopher* (Courtois, 1994). Diferentemente do *Veronica*, que se destacava pela indexação em larga escala de títulos de menus distribuídos globalmente, o *Jughead* apresentava uma abordagem mais restrita, voltada à realização de buscas em servidores *Gopher* individuais (Courtois, 1994). Nesse contexto, a ferramenta operava como um mecanismo de exploração hierárquica local, permitindo a recuperação eficiente de informações dentro de um único domínio (Davidoff, 1994). Tal característica atendia às demandas de usuários

e administradores que necessitavam de maior rapidez e precisão na localização de conteúdos específicos, sem a dependência de índices distribuídos de maior complexidade.

Do ponto de vista técnico, o *Jughead* varria a hierarquia de menus de um servidor *Gopher* e construía uma base de dados local de títulos e descrições, permitindo que consultas textuais fossem respondidas de forma quase imediata (Courtois, 1994). O software era leve, de fácil instalação e não exigia grandes recursos computacionais, o que o tornou bastante popular em universidades e instituições de pesquisa (Courtois, 1994). Assim como o *Veronica*, o *Jughead* integrava-se à lógica de menus do *Gopher*, retornando os resultados em forma de listas navegáveis, preservando a experiência de uso do *Gopherspace* (Anklesaria *et al*., 1993).

A principal contribuição do *Jughead* foi oferecer um mecanismo de busca mais próximo do usuário e de baixo custo, ampliando a flexibilidade do ecossistema *Gopher*. Enquanto o *Veronica* respondia à necessidade de uma busca global, o *Jughead* supria demandas locais, mostrando que a Recuperação da Informação na internet exigia soluções em múltiplas escalas (Courtois, 1994). Apesar de ter perdido relevância com a ascensão da World Wide Web e de mecanismos mais avançados de indexação, como o *WebCrawler*, o *Jughead* representou um elo importante na evolução dos motores de busca ao demonstrar a viabilidade de ferramentas segmentadas e contextuais (Courtois, 1994).

4.4.1 Aplicações Práticas do Jughead

Uma aplicação prática do *Jughead* ocorreu em universidades e instituições que utilizavam servidores *Gopher* para organizar catálogos de cursos, regulamentos internos, arquivos de pesquisa e materiais administrativos. Ao contrário do *Veronica*, que buscava resultados em escala global, o *Jughead* era empregado para buscas locais, permitindo que usuários encontrassem documentos em um servidor específico (Courtois, 1994). Essa característica foi especialmente útil em ambientes institucionais, onde grandes volumes de dados precisavam ser acessados por estudantes, professores e funcionários sem a necessidade de recorrer a sistemas externos (Davidoff, 1994). O acrônimo "Jughead"foi posteriormente definido como *Jonzy's Universal Gopher Hierarchy Excavation And Display*, conforme documentado em relatório técnico do Argonne National Laboratory, apresentado em workshop sobre Internet em 1994 (Davidoff, 1994).

4.4.2 Vantagens e Desvantagens

O *Jughead* surgiu em um contexto no qual a demanda por ferramentas de recuperação de informação crescia de forma acelerada, especialmente com a expansão do *Gopherspace* (Anklesaria *et al.*, 1993). Ele representou um avanço no sentido de oferecer maior autonomia aos administradores e usuários de servidores *Gopher*, embora também apresentasse limitações frente a soluções mais abrangentes (Courtois, 1994).

4.4.2.1 Vantagens do Jughead na Época

De acordo com a literatura técnica da época, algumas vantagens do *Jughead* incluíam:

- Simplicidade de instalação e manutenção, tornando-o acessível mesmo para pequenas instituições (Davidoff, 1994);
- Baixo consumo de recursos computacionais, ideal para a infraestrutura limitada do início da década de 1990 (Courtois, 1994);
- Rapidez na recuperação de informações locais, já que operava sobre índices restritos a um único servidor (Anklesaria *et al.*, 1993);
- Complementaridade ao *Veronica*, oferecendo uma alternativa de busca em escala reduzida (Courtois, 1994).

4.4.2.2 Desvantagens do Jughead na Época

Por outro lado, as limitações do *Jughead* também foram amplamente documentadas:

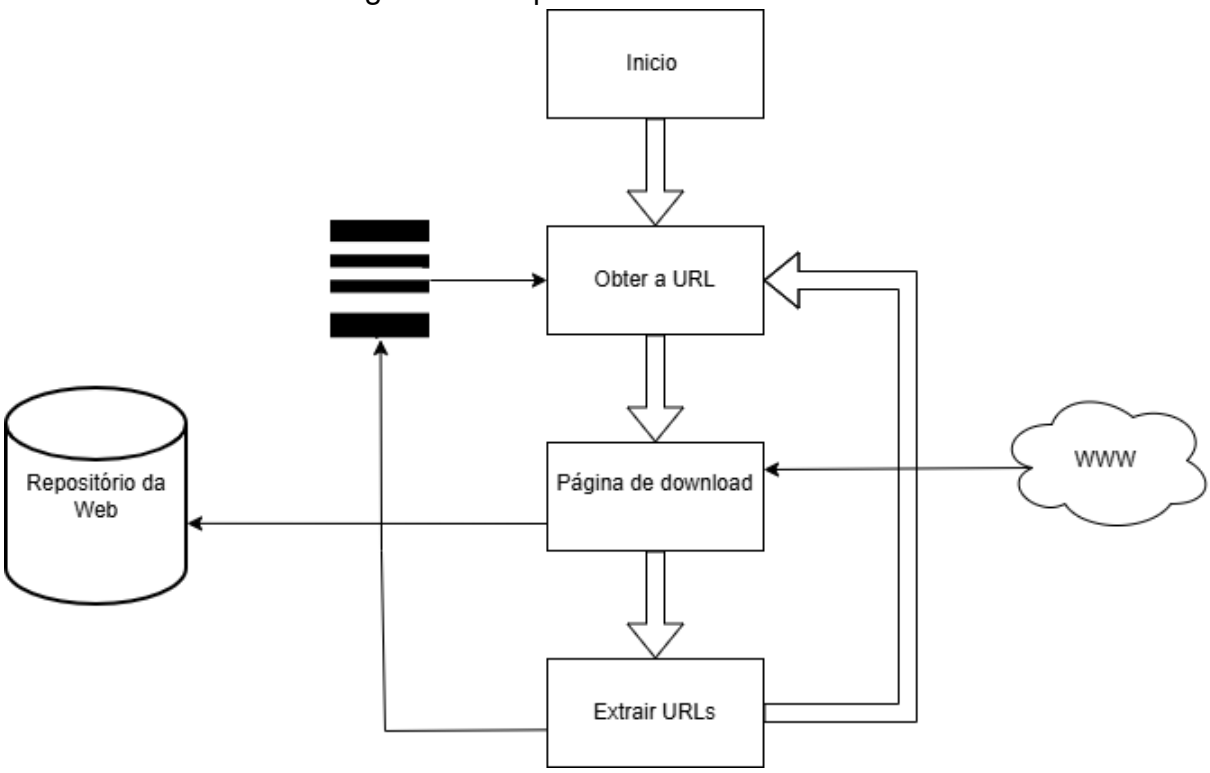
- Restrição ao âmbito local, sem capacidade de integrar resultados de múltiplos servidores, como fazia o *Veronica* (Courtois, 1994);
- Dependência de administradores locais para manutenção e atualização da base de dados (Davidoff, 1994);
- Ausência de funcionalidades mais sofisticadas de indexação e ranqueamento, que começaram a ser introduzidas em motores de busca posteriores (Courtois, 1994);
- Perda de relevância com o crescimento exponencial da World Wide Web e de buscadores globais baseados em hipertexto, que tornaram o Gopher obsoleto (Courtois, 1994).

4.5 Surgimento e Mecanismo de Funcionamento do Web Crawler

O *Web Crawler*, desenvolvido em 1994 por Brian Pinkerton, na Universidade de Washington, foi o primeiro mecanismo de busca de texto completo amplamente disponível na *web* (Kausar *et al.*, 2013). Diferentemente de sistemas anteriores, ele permitia que os usuários pesquisassem no conteúdo integral das páginas, em vez de depender apenas de palavra-chave ou descrições fornecidas manualmente, o que reduzia a possibilidade de resultados confusos, tornando, assim, as buscas mais eficazes (Kausar *et al.*, 2013).

O funcionamento de um *Web Crawler* começa com um conjunto inicial de URLs, conhecido como *seed* URLs (PALHARINI *et al.*, 2023). Ele faz o *download* das páginas da *web* correspondentes a essas *seed* URLs e extrai novos *links* presentes nas páginas baixadas (Kausar *et al.*, 2013). As páginas recuperadas são armazenadas e indexadas em uma área de armazenamento, de forma que, com a ajuda desses índices, possam ser recuperadas posteriormente, sempre que necessário (PALHARINI *et al.*, 2023). As URLs extraídas da página baixada são verificadas para saber se seus documentos associados já foram baixados ou não. Se ainda não tiverem sido baixadas, as URLs são novamente atribuídas aos *web crawlers* para novos *downloads* (Kausar *et al.*, 2013). Este processo se repete até que não haja mais URLs pendentes para baixar. Milhões de páginas são baixadas por dia por um *crawler* para completar o objetivo. A Figura 5 ilustra os processos de *crawling*.

Figura 5 – Arquitetura do Web Crawler



Fonte: Adaptado de: (Kausar *et al.*, 2013).

Segundo (Kausar *et al.*, 2013), o funcionamento de um *web crawler*, que consiste no processo de *crawling*, pode ser discutido da seguinte forma:

a) Selecionar uma URL inicial (*seed* URL) ou várias URLs;
b) Adicioná-las à fronteira;
c) Em seguida, escolher a URL da fronteira;
d) Buscar a página *web* correspondente àquela URL;
e) Analisar (*parsing*) a página *web* para encontrar novos *links* de URL;
f) Adicionar todas as URLs recém-encontradas à fronteira;
g) Voltar ao passo "b"e repetir até que a fronteira esteja vazia.

Assim, um *web crawler* continuará inserindo recursivamente novas URLs no repositório do banco de dados do mecanismo de busca. Dessa forma, pode-se ver que a principal função de um *web crawler* é inserir novos *links* na fronteira e escolher uma URL inédita da fronteira para processamento posterior após cada passo recursivo.

4.6 Funcionamento do Motor de Busca AltaVista

O motor de busca AltaVista, lançado em dezembro de 1995 pela Digital Equipment Corporation (DEC), tornou-se um marco tecnológico no processo de recu-

peração de informações. Sua construção foi orientada para garantir alto desempenho em indexação, escalabilidade e velocidade na apresentação de resultados, uma necessidade crescente devido à rápida expansão da web na década de 1990 (Silverstein *et al.*, 1999).

O AltaVista foi projetado para operar com grande capacidade computacional, utilizando servidores baseados no processador Alpha 8400 (Monier, 1997). Essa infraestrutura permitia lidar com milhões de documentos e consultas diárias, tornando-o um dos motores de busca mais avançados de sua época (Monier, 1997).

4.6.1 Arquitetura Funcional do AltaVista

O funcionamento do mecanismo de busca era estruturado de forma modular, conforme ilustrado na Tabela 8, com etapas específicas para coleta, tratamento e recuperação da informação (Silverstein *et al.*, 1999).

Tabela 8 – Modos de Consulta do AltaVista

Componente	**Descrição**
Crawler Scooter	Responsável por explorar páginas da web automaticamente, seguindo hiperlinks e copiando documentos para o servidor. Operava continuamente para manter o índice atualizado.
Processador/Indexador	Realizava a análise dos documentos coletados, extraindo texto relevante, identificando termos, construindo índices invertidos e armazenando metadados.
Banco de Dados Indexado	Repositório de documentos já analisados com estrutura otimizada de busca, garantindo resposta rápida às consultas.
Mecanismo de Consulta	Interpretava consultas dos usuários, realizava buscas no índice e retornava resultados ordenados por relevância.
Interface Web	Permitia que os usuários interagissem com o sistema através do endereço www.altavista.com.

Fonte: Elaborado pelo próprio autor, 2025.

4.6.2 Rastreamento e Coleta de Páginas Web

O processo de rastreamento era realizado pelo *Scooter*, um dos primeiros *web crawlers* de alto desempenho (Monier, 1997). Seu funcionamento, de acordo com (Monier, 1997), era um fluxo operacional estruturado, composto pelas seguintes etapas:

a) Acessar um conjunto inicial de URLs conhecidas (*seeds*);

b) Fazer o download completo dos documentos;
c) Extrair e catalogar os hiperlinks encontrados;
d) Adicionar novos links à fila de rastreamento;
e) Repetir o processo continuamente.

Esse mecanismo possibilitava uma grande cobertura da *web*, chegando a 6 milhões de páginas indexadas por dia ainda nos anos iniciais de operação (Monier, 1997).

4.6.3 Indexação e Organização da Informação

Após a coleta, os documentos eram enviados ao módulo indexador, no qual o conteúdo era processado por diversas técnicas, como:

a) Filtragem de *tags* HTML;
b) Normalização textual (BAEZA-YATES *et al.*, 1999);
c) Tokenização de palavras-chave (BAEZA-YATES *et al.*, 1999);
d) Armazenamento em índices invertidos (Robertson, 1977).

O índice invertido permitia que o sistema identificasse rapidamente em quais documentos determinados termos apareciam, aumentando o desempenho da busca (BAEZA-YATES *et al.*, 1999).

A Tabela 9 apresenta as principais etapas de indexação (Brin *et al.*, 1998).

Tabela 9 – Etapas do Processo de Indexação do AltaVista

Etapa	**Descrição**
Extração de conteúdo	Remoção de elementos estruturais, como código HTML, para obtenção do texto puro;
Análise linguística	Identificação de termos relevantes para busca, desconsiderando palavras com baixo valor semântico;
Criação do índice invertido	Associação de cada termo aos documentos onde aparecem, viabilizando recuperação eficiente;
Armazenamento segmentado	Organização otimizada do índice para permitir alta velocidade e múltiplas consultas simultâneas.

Fonte: Elaborado pelo próprio autor, 2025.

4.6.4 Consultas e Recuperação dos Resultados

O AltaVista empregava um modelo de recuperação baseado em busca booleana ponderada, um sistema que permitia aos usuários estruturar consultas de maneira flexível (ROUSSEAU, 1998). Esse modelo era dividido em dois modos principais de operação, conforme detalhado por (ROUSSEAU, 1998):

4.6.4.1 Modos de Operação

- **Busca Simples:** O modo padrão em que o sistema combinava automaticamente os termos inseridos utilizando o operador OR. Isso maximizava o número de documentos recuperados (Silverstein *et al.*, 1999);
- **Busca Avançada:** Permitia que o usuário empregasse explicitamente os operadores booleanos e de proximidade, possibilitando maior controle e precisão sobre os resultados (Silverstein *et al.*, 1999).

4.6.4.2 Operadores de Consulta e Lógica Booleana

Na Busca Avançada, o AltaVista utilizava uma sintaxe correspondente à lógica booleana tradicional. Os operadores mais importantes estão listados na Tabela 10 (Silverstein *et al.*, 1999).

Tabela 10 – Operadores de Consulta do AltaVista

Componente da Consulta	**Descrição**
`+termo` (AND)	O documento **obrigatoriamente** deve conter o termo solicitado para ser recuperado;
`-termo` (NOT)	O documento que contém o termo é **excluído** da lista de resultados;
`"frase exata"`	Exige que os termos estejam **adjacentes** e na ordem exata em que foram digitados.

Fonte: Elaborado pelo autor (2025).

4.6.4.3 Ordenação e Relevância

A recuperação não se limitava apenas à presença ou ausência dos termos. Após a filtragem booleana, os resultados eram exibidos para o usuário ordenados por critérios de relevância (Silverstein *et al.*, 1999). Essa pontuação era determinada por fatores como a frequência e a posição dos termos dentro de cada documento, garantindo que os resultados mais pertinentes aparecessem no topo da lista (Silverstein *et al.*, 1999).

4.6.5 Inovações e Contribuições Tecnológicas

O *AltaVista* foi responsável por avanços significativos para os sistemas de busca modernos, destacando-se:

- Alta escalabilidade na indexação de documentos (Monier, 1997);

- Disponibilidade de consultas avançadas com suporte a operadores booleanos (Silverstein *et al.*, 1999).

Essas contribuições ajudaram a consolidar a recuperação de informação na *web* como um campo essencial na computação e comunicação digital.

4.6.6 Cadê?

O Cadê? foi um marco fundamental e pioneiro na história da Internet no Brasil, sendo lançado em setembro de 1995 pelos estudantes Gustavo Viberti e Fábio Oliveira com o objetivo inicial de catalogar manualmente os sites nacionais em um diretório funcional, preenchendo a lacuna de ferramentas de busca focadas no conteúdo em português (Gomes *et al.*, 2001). Rapidamente tornou-se o principal *gateway* de acesso à informação *online* no país, expandindo-se para indexar diversos tipos de conteúdo, consolidando-se como uma referência de mercado, até ser adquirido pelo Yahoo! Brasil em 2003 e, posteriormente, descontinuado, mas deixando o legado de ter sido a primeira grande iniciativa brasileira a organizar e popularizar o acesso digital à informação no país (Gomes *et al.*, 2001).

4.7 Google: A Revolução na Recuperação de Informação

O Google foi fundado por Larry Page e Sergey Brin enquanto ambos eram estudantes de pós-graduação na Universidade de Stanford. A proposta inicial do projeto consistia em usar a estrutura de ligações (links) entre páginas para estimar a importância e relevância dos documentos na *web*, uma ideia que resultou no algoritmo denominado *PageRank* (Moraes Braz *et al.*, 2012). Em 1998, a empresa foi formalmente constituída, e rapidamente o sistema de busca desenvolvido por Page e Brin passou a ser reconhecido por apresentar resultados mais relevantes que os dos mecanismos de busca dominantes à época (Oliveira *et al.*, 2016).

4.7.1 Arquitetura geral de um mecanismo de busca moderno

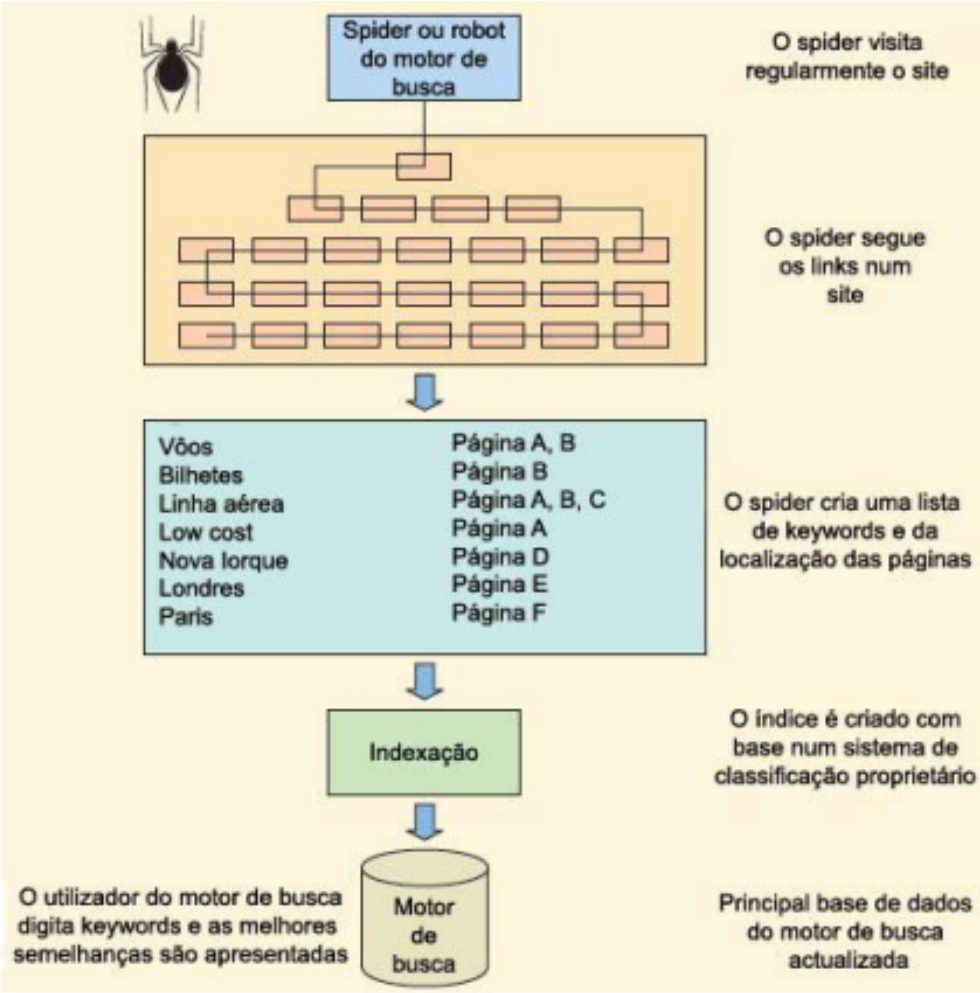
De forma resumida, o funcionamento de um mecanismo de busca como o Google pode ser dividido em três etapas principais:

a) ***Crawling*:** programas automatizados conhecidos como *web crawlers* que percorrem a *web* seguindo links e descobrindo conteúdo novo ou atualizado (PALHARINI *et al.*, 2023);
b) **Indexação:** as informações que são recolhidas são armazenadas e indexadas na base de dados (Caldeira, 2015);
c) **Ranking e recuperação:** quando um usuário submete uma consulta,

o sistema a compara com o índice e ordena os possíveis resultados segundo um conjunto de sinais e modelos que estimam relevância e qualidade (Ferneda, 2003).

A Figura 6 apresenta um modelo visual que ilustra as etapas principais do funcionamento de um mecanismo de busca.

Figura 6 – Arquitetura de um mecanismo de busca



Fonte: (ResearchGate, s.d.).

4.7.2 O algoritmo PageRank: ideia e formulação

Uma das principais contribuições algorítmicas dos primórdios do Google é o *PageRank*, que quantifica a importância de uma página com base na estrutura de ligações da *web*, considerando que páginas mais relevantes tendem a receber mais referências de outras páginas (Caldeira, 2015).

Intuitivamente, o PageRank pode ser interpretado por meio do modelo do *random surfer* (navegador aleatório), no qual um usuário percorre a Web seguindo

links de forma probabilística. Nesse modelo, páginas que são mais frequentemente visitadas por esse navegador aleatório possuem maior importância.

Em sua forma matemática, o PageRank atribui a cada página i um valor $PR(i)$, definido recursivamente por:

$$PR(i) = \frac{1-d}{N} + d \sum_{j \in B_i} \frac{PR(j)}{L(j)}$$

onde:

- N é o número total de páginas no grafo da Web;
- d é o fator de amortecimento (*damping factor*), geralmente definido como 0.85, representando a probabilidade de o usuário continuar navegando por links;
- B_i é o conjunto de páginas que apontam para i;
- $L(j)$ é o número de links de saída da página j.

O fator de amortecimento d também desempenha um papel fundamental na garantia da convergência do algoritmo, ao permitir que o modelo considere a possibilidade de o usuário saltar para qualquer página aleatoriamente, evitando problemas associados a ciclos e componentes desconexos no grafo.

Do ponto de vista matemático, o *PageRank* corresponde ao autovetor dominante de uma matriz de transição estocástica associada ao grafo de links da Web (Haveliwala, 1999). Essa matriz define uma cadeia de Markov, cuja distribuição estacionária representa os valores de *PageRank* das páginas.

Computacionalmente, o vetor de *PageRank* é obtido por meio de métodos iterativos, como o método da potência (*power iteration*), que atualiza sucessivamente os valores até que a convergência seja atingida (Haveliwala, 1999). Esse procedimento é eficiente e escalável, permitindo sua aplicação em grafos de grande dimensão, como a Web.

4.7.3 Como o PageRank se integra ao processo de ranking

O *PageRank*, proposto originalmente por Brin e Page, é um algoritmo baseado na estrutura de links da *web* que atribui maior importância às páginas que recebem links de outros sites considerados relevantes (Brin *et al.*, 1998). Essa abordagem permitiu captar uma noção de autoridade, associada à popularidade medida por referências externas (Page *et al.*, 1999). No entanto, embora fundamental em sua época, o *PageRank* rapidamente se mostrou insuficiente como critério único na ordenação dos resultados de pesquisa (Souza, 2006).

Outro conjunto de fatores determinantes no ranking advém do comportamento dos usuários. De acordo com (Glick *et al.*, 2014), os cliques, padrões de na-

vegação e preferência dos usuários são incorporados como sinais de envolvimento, permitindo que o motor de busca ajuste dinamicamente a importância relativa de cada página. Dessa forma, a experiência real do usuário torna-se parte do modelo de decisão.

Além disso, variantes do *PageRank* tradicional foram desenvolvidas para superar a distribuição uniforme de importância, como demonstrado no Weighted *PageRank* (WPR) proposto por Xing e Ghorbani em 2004 (WEIGHTED. . . , 2023). Diferentemente do algoritmo original, o WPR introduz pesos diferenciados aos *links* a partir de métricas de popularidade, levando em conta a relevância tanto dos *links* de entrada *in-links* quanto dos de saída *out-links* (Naamha *et al.*, 2024). Em vez de dividir o ranking igualmente entre todas as páginas conectadas, essa adaptação atribui valores proporcionais à importância relativa de cada link, revelando a evolução do algoritmo para uma abordagem mais holística e contextual (Naamha *et al.*, 2024).

Por fim, os motores de busca modernos também utilizam sinais contextuais, tais como localização geográfica, tipo de dispositivo utilizado e horário de acesso, que influenciam a personalização dos resultados (Milvus Research, 2023). Esses sinais completam um sistema de classificação complexo, cujo objetivo é apresentar ao usuário resultados cada vez mais precisos e ajustados à sua intenção.

4.7.4 Evolução dos algoritmos de ranking (principais marcos)

Desde o lançamento inicial, o Google realizou várias atualizações e reengenharias em seus algoritmos de rankeamento, aprimorando constantemente sua capacidade de oferecer resultados mais relevantes e contextualizados aos usuários. Alguns dos marcos mais destacados na literatura incluem:

- **Panda (2011):** introduzido com foco em penalizar conteúdo de baixa qualidade, raso ou duplicado, o algoritmo Panda representou um marco importante na promoção de conteúdo original e relevante (Google Search Central Blog, 2011).
- **Penguin (2012):** projetado para combater práticas de manipulação de links, como o *link spam*, o Penguin reforçou a importância da qualidade dos backlinks em detrimento da quantidade (Google Search Central Blog, 2016).
- **Hummingbird (2013):** uma reescrita profunda da infraestrutura do motor de busca, com o objetivo de melhorar o entendimento de consultas naturais e da intenção por trás das buscas, marcando a transição para uma compreensão semântica mais ampla (Sullivan, 2013).
- **RankBrain (2015):** ponto de virada no uso de *machine learning*, o *RankBrain* passou a interpretar consultas inéditas e a ajustar automatica-

mente os sinais de relevância, indo além de regras estáticas e introduzindo adaptação inteligente aos resultados (Sullivan, 2018).

- **BERT (2019):** com base em arquiteturas de transformadores, o BERT trouxe melhorias decisivas na compreensão do contexto de palavras em consultas, refinando a análise semântica e a relação entre termos (Singh, 2021).
- **Multitask Unified Model (MUM) e atualizações recentes:** o MUM expande o uso de IA no sistema de buscas, sendo multimodal e capaz de processar informações a partir de texto, imagens e outros formatos, além de operar em múltiplos idiomas (Schwartz, 2022).

Esses marcos demonstram a evolução de algoritmos baseados em regras fixas e sinais tradicionais, para arquiteturas que integram modelos estatísticos complexos, aprendizado profundo e compreensão semântica avançada. Trata-se de uma tendência em que os mecanismos de busca se tornam cada vez mais capazes de interpretar linguagem natural e intenção de busca, garantindo, assim, maior precisão e personalização nos resultados.

4.7.5 Desafios, controvérsias e considerações éticas

Com grande poder de filtragem e influência sobre a visibilidade da informação, o Google enfrentou (e enfrenta) desafios importantes:

- **Transparência:** a opacidade dos modelos de ranking levanta questões sobre responsabilidade e viés (Pasquale, 2015; Barredo Arrieta *et al.*, 2020).
- **Privacidade:** sinais de personalização dependem de dados de usuários; há tensão entre relevância e privacidade (Acar *et al.*, 2014; Erlingsson *et al.*, 2014).
- **Monopólio e competição:** o domínio de mercado gerou debates regulatórios e antitruste em várias jurisdições (Yun, 2018; Varian, 2019).

4.8 Redes Sociais como Motores de Busca ?

A proliferação das plataformas digitais de interação social transformou a lógica de acesso à informação, a ponto de configurar algumas redes sociais como verdadeiros motores de busca informacionais paralelos aos mecanismos tradicionais de busca *web*. Antes centradas apenas na interconexão de pessoas e no compartilhamento de conteúdo, essas redes constituem agora ambientes em que usuários buscam, filtram e consomem informação de modo direto, deslocando, em certa medida, o locus da “busca” para o *feed* social.

4.8.1 Mudança de paradigma no comportamento de busca

Estudos com usuários brasileiros apontam que as redes sociais são cada vez mais utilizadas como fontes de informação. Por exemplo, a pesquisa da Kaspersky, em parceria com a Corpa, verificou que 71% dos usuários brasileiros entre 20 e 65 anos recorreram a redes sociais como fonte de informação em meados de 2023 (Kaspersky, 2023). Outro trabalho evidencia que, entre estudantes da Universidade Federal de Pernambuco (UFPE), as redes sociais como Facebook, WhatsApp, YouTube, Instagram e Twitter foram apontadas como “ferramenta de busca informacional” para temas diversos, inclusive acadêmicos (CAETANO, 2018). Esses achados indicam que há uma reelaboração do papel informacional dessas plataformas: não mais somente como locais de circulação de informação, mas como ambientes de consultas, investigação e descoberta.

4.8.2 Características que permitem o funcionamento como motor de busca

Para que as redes sociais operem como motores de busca, é preciso considerar algumas de suas características específicas, que as diferenciam dos mecanismos tradicionais (como os buscadores web):

a) **Feed personalizado e curadoria social**: o usuário é exposto a conteúdos recomendados com base em seus interesses, comportamentos e rede de contatos, o que facilita encontrar informação relevante sem iniciar uma consulta explícita (Marteleto, 2007);
b) **Interação social e reputação**: a presença de comentários, compartilhamentos e “curtidas” cria um sistema de valoração que orienta o usuário para conteúdos mais reputados ou populares, semelhantemente ao ranking em motores de busca (Marteleto, 2007; Pennini *et al.*, 2018);
c) **Multiformatos e multimídia**: as redes comportam textos, vídeos, imagens e áudios, ampliando o escopo da busca para formatos não tradicionais (Oliveira Souza *et al.*, 2021).
d) **Acesso móvel e ubiquidade**: o uso intenso via dispositivos móveis torna as redes sociais ambientes acessíveis e imediatos, favorecendo a busca informal e espontânea (Kaspersky, 2023);
e) **Ambientes híbridos entre busca e socialização**: a consulta à informação ocorre embebida no contexto da interação social, diferindo da busca neutra típica dos mecanismos convencionais (Marteleto, 2007).

4.8.3 Potencialidades e desafios

As redes sociais como ferramentas de busca trazem importantes potencialidades para o cenário informacional: democratização do acesso à informação, agilidade na descoberta de informações emergentes e a possibilidade de conectar a consulta à rede social de apoio e validação.

Entretanto, emergem desafios que demandam atenção crítica: a confiabilidade das informações, conforme indicou o estudo da UFPE,pois, embora os estudantes utilizem redes sociais para buscar informação, há dúvidas quanto à sua validação como fonte de pesquisa (CAETANO, 2018); o viés de popularidade e as bolhas de informação; a dificuldade em recuperar a historicidade da informação; e a falta de padronização e indexação formal dos conteúdos compartilhados.

4.8.4 Implicações para a Ciência da Informação e para práticas de busca

No contexto da Ciência da Informação, a reelaboração das redes sociais como ambientes de busca exige repensar conceitos como “fonte de informação”, “busca eficiente” e “comportamento informacional”. Segundo (Marteleto, 2007) já destacava a rede social como um contexto em que a informação se articula com a estrutura coletiva de interação considerando redes sociais como “modos de emprego da informação”. Além disso, o estudo de Souza, Neves e Mello (Souza *et al.*, 2024) enfatiza como essas plataformas constituem ecossistemas de acesso à informação e de políticas de internet. Logo, é relevante para pesquisadores e profissionais considerar que a busca informacional em redes não segue os mesmos modelos de motores tradicionais e requer adaptações metodológicas (como análise de *feeds*, métricas de engajamento e interação social) para mapear como a informação é localizada e consumida.

5 CONCLUSÃO

O presente trabalho constituiu-se em uma RSL acerca dos motores de busca, com foco nos processos de recuperação de informações. Buscou-se compreender como esses sistemas foram concebidos, evoluíram e se consolidaram ao longo do tempo, apresentando seu desenvolvimento histórico por meio de uma linha temporal que evidencia os principais marcos tecnológicos da área. Além disso, como objetivo secundário, procurou-se analisar a evolução das funcionalidades, a relevância tecnológica e o impacto desses mecanismos na organização, no acesso e na disponibilização da informação, considerando sua influência tanto no âmbito acadêmico quanto no cotidiano dos usuários da web.

Como resultado da investigação conduzida, e em consonância com os objetivos delineados, foi possível elaborar uma linha temporal que contempla os principais motores de busca mencionados na literatura especializada. Essa trajetória histórica permitiu identificar sistemas que marcaram profundamente o desenvolvimento da web, cada qual apresentando características particulares que contribuíram para a consolidação da RI como campo estratégico na ciência da computação. A análise desses sistemas possibilitou compreender como avanços tecnológicos, mudanças de paradigma e necessidades crescentes de organização da informação impulsionaram a criação de novas soluções ao longo dos anos.

Entre os motores de busca analisados, constatou-se que o Archie foi um marco inicial ao viabilizar a localização de arquivos em servidores FTP, introduzindo novas possibilidades de recuperação da informação em um período em que a internet ainda era incipiente. Dando continuidade à evolução histórica, identificou-se que o Gopher se destacou como um dos primeiros sistemas estruturados para a organização e navegação hierárquica de conteúdos, oferecendo ao usuário uma perspectiva mais sistemática de acesso a documentos e serviços. Complementarmente, o Jughead emergiu como uma solução voltada para a indexação de menus Gopher, ampliando sua capacidade de busca; enquanto o Veronica apresentou um avanço significativo ao permitir buscas globais nos títulos desses menus, tornando o processo mais dinâmico e abrangente.

Com o amadurecimento da web e o aumento exponencial de informações online, surgiram motores de busca mais sofisticados, entre os quais, se destaca o Google, que revolucionou o cenário ao ampliar drasticamente o alcance, a precisão e a capacidade de indexação de conteúdos. Seu algoritmo de ranqueamento, baseado na relevância, trouxe novos parâmetros para avaliação da informação, consolidando-o como um dos mecanismos mais influentes da história da internet. Além disso, observa-se que, com a ascensão da geração Z e a intensificação do uso das redes sociais, plataformas como TikTok, Instagram e YouTube passaram a desempenhar um

papel crescente como ferramentas de busca, evidenciando uma mudança nos hábitos informacionais e no modo como jovens usuários interagem com conteúdos digitais.

Por fim, para trabalhos futuros, recomenda-se a realização de pesquisas de campo que investiguem, de forma aprofundada, como estudantes efetuam suas buscas atualmente. Tais estudos poderiam analisar suas estratégias de pesquisa, as dificuldades enfrentadas, as ferramentas que utilizam com maior frequência e o nível de conhecimento que possuem sobre o funcionamento dos motores de busca tradicionais e alternativos. Investigações desse tipo podem gerar insights relevantes para a formulação de práticas educacionais mais eficazes, direcionadas ao fortalecimento da competência informacional, ao desenvolvimento do pensamento crítico e ao uso consciente e qualificado das tecnologias voltadas à recuperação da informação. Dessa forma, amplia-se a compreensão sobre o comportamento dos usuários contribuindo para a formação de indivíduos mais preparados para lidar com o crescente volume de dados disponível no ecossistema informacional contemporâneo.

REFERÊNCIAS

ACAR, G. *et al.* The Web Never Forgets: Persistent Tracking Mechanisms in the Wild. **Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security**, p. 674–689, 2014.

ANKLESARIA, F. *et al.* **The Internet Gopher Protocol: a distributed document search and retrieval protocol**. 1993. p. 16. RFC 1436. Disponível em: https://www.rfc-editor.org/rfc/rfc1436.

ARAÚJO, C. A. Á. **O que é ciência da informação**. Belo Horizonte: KMA, 2018. p. 126. ISBN 978-85-92728-06-9. Disponível em: https://teste.eci.ufmg.br/wp-content/uploads/2024/03/O-QUE-E-CIENCIA-DA-INFORMACAO.pdf.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: Addison-Wesley, 1999. Disponível em: https://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf.

BARREDO ARRIETA, A. *et al.* Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82–115, 2020.

BERNERS-LEE, T. *et al.* The World-Wide Web. **Communications of the ACM**, v. 37, n. 8, p. 76–82, 1994.

BOS, B. **Representation Protocols**. 1997. Disponível em: https://www.w3.org/People/Bos/PROSA/rep-protocols.html.

BRIN, S.; PAGE, L. The anatomy of a large-scale hypertextual Web search engine. **Computer networks and ISDN systems**, Elsevier, v. 30, n. 1-7, p. 107–117, 1998.

CAETANO, D. d. C. **O comportamento informacional no uso de redes sociais virtuais como fonte de informação**. Recife, 2018. Trabalho de Conclusão de Curso (Graduação em Biblioteconomia) – Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Departamento de Ciência da Informação. Disponível em: https://repositorio.ufpe.br/handle/123456789/30680.

CALDEIRA, F. H. O mecanismo de busca do Google e a relevância na relação sistema-usuário. **Letrônica: Revista Digital do Programa de Pós-Graduação em**

Letras da PUCRS, Porto Alegre, v. 8, n. 1, p. 91–106, 2015. Acesso em: 28 out. 2025. Licença Creative Commons Atribuição 4.0 Internacional. ISSN 1984-4301. DOI: 10.15448/1984-4301.2015.1.19616. Disponível em: https://revistaseletronicas.pucrs.br/letronica/article/view/19616.

CARDILLO, A. **Quantos sites existem na internet? (2025)**. 2025. Exploding Topics. Acesso em: 7 jun. 2025. Disponível em: https://explodingtopics.com/blog/how-many-websites-on-the-internet.

CENDÓN, B. V. Ferramentas de busca na Web. **Ciência da Informação**, Associação Brasileira de Tecnologia em Saúde, v. 30, n. 1, p. 39–49, 2001. Disponível em: https://www.scielo.br/j/ci/a/WdYRz6LmQbBD5ZWnKTfHKkm/?format=pdflang=pt. Acesso em: 7 jun. 2025.

CIOLEK, T. M. **Asian Studies and the WWW: a Quick Stocktaking at the Cusp of Two Millennia**. Apresentado em 15–18 May 1998. Pacific Neighbourhood Consortium Annual Meeting. 1998. Disponível em: https://pnclink.org/annual/annual1998/1998pdf/ciolek.htm.

COURTOIS, M. P. How to find information using Internet Gophers. **Online**, v. 18, n. 6, p. 14–25, 1994. Disponível em: https://portaildoc-intd.cnam.fr/Record.htm?idlist=1&record=19139464124919576469.

DAVIDOFF, G. **Using Gopher on the Internet**. 1994. p. 29. Apresentado no workshop da Suburban Library, Burr Ridge, IL, 26 abr. 1994. Disponível em: https://digital.library.unt.edu/ark:/67531/metadc1312069/.

DEVLIN, J. *et al.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *In:* ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. PROCEEDINGS of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. p. 4171–4186. DOI: 10.18653/v1/N19-1423. Disponível em: https://aclanthology.org/N19-1423/.

DOMINGUES, A. S. **MOTOR DE BUSCA PARA INTEGRAÇÃO DE DADOS SOBRE SAÚDE NA WEB**. MOSSORÓ, 2019. Disponível em: https://repositorio.ufersa.edu.br/server/api/core/bitstreams/272f6199-07c9-436d-9364-c027b50695da/content. Acesso em: 7 jun. 2025.

EIBEN, A.; SMITH, J. **Introduction To Evolutionary Computing**. 01/2003. v. 45. ISBN 978-3-642-07285-7. DOI: 10.1007/978-3-662-05094-1.

ERLINGSSON, U.; PIHUR, V.; KOROLOVA, A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. **Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security**, p. 1054–1067, 2014.

ESCANDELL-POVEDA, R.; IGLESIAS-GARCÍA, M.; PAPÍ-GÁLVEZ, N. From Memex to Google: The origin and evolution of search engines*. In:* 2022. p. 47–66. ISBN 978-84-120239-9-2. DOI: 10.3145/indocs.2022.4.

FERNEDA, E. **Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para Ciência da Informação**. 2003. Tese (Doutorado em Ciências da Comunicação) – Escola de Comunicação e Artes da Universidade de São Paulo, São Paulo. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/Tese.pdf. Acesso em: 7 jun. 2025.

FERNEDA, E.; DIAS, G. A. A lógica fuzzy aplicada à recuperação de informação. **InterScientia**, InterScientia, João Pessoa, v. 1, n. 1, p. 51–65, 2013. jan./abr. 2013.

GIARRATANO, J. C.; RILEY, G. **Expert systems: principles and programming**. 4th. Boston, MA: Thomson Course Technology, 2005. p. 842. Acompanha 1 CD-ROM. ISBN 0534384471.

GLICK, M. *et al.* How Does Ranking Affect User Choice in Online Search? **Review of Industrial Organization**, v. 45, p. 99–119, 2014. DOI: 10.1007/s11151-014-9435-y.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. Reading, MA: Addison-Wesley, 1989.

GOMES, E. A. *et al.* **Fidus: uma ferramenta para busca de informações personalizadas na web**. 2001. Dissertação (Mestrado) – Universidade Federal da Paraíba – Campus II, Paraíba.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. http://www.deeplearningbook.org.

GOOGLE SEARCH CENTRAL BLOG. **More guidance on building high-quality sites**. 2011. https://developers.google.com/search/blog/2011/05/more-guidance-on-building-high-quality?hl=pt-br. Acesso em: 4 nov. 2025.

GOOGLE SEARCH CENTRAL BLOG. **O Penguin agora faz parte do nosso algoritmo principal**. 2016. https://developers.google.com/search/blog/2016/09/penguin-is-now-part-of-our-core?hl=pt-br. Acesso em: 4 nov. 2025.

HAVELIWALA, T. H. **Efficient Computation of PageRank**. Stanford, CA, 1999. p. 15. Acesso em: 28 abr. 2026. Disponível em: http://ilpubs.stanford.edu:8090/386/.

HAYKIN, S. **Neural Networks and Learning Machines**. 3. ed.: Pearson, 2009.

IETF RFC EDITOR. **RFC 1689: A Status Report on Networked Information Retrieval: Tools and Groups**. 1994. RFC.

JARA MORANTE, W. **Diseño e implementación de un sistema de información para la ESPOL "SI-ESPOL" basado en la aplicación Gopher de la Universidad de Minnesota**. 1994. Tesis de grado – Escuela Superior Politécnica del Litoral (ESPOL), Guayaquil.

KASPERSKY. **Redes sociais são usadas por 71% dos brasileiros como fonte de informação, mostra pesquisa**. Kaspersky Brasil. 2023. Disponível em: https://www.kaspersky.com.br/about/press-releases/redes-sociais-sao-usadas-por-71-dos-brasileiros-como-fonte-de-informacao-mostra-kaspersky.

KAUSAR, M. A.; DHAKA, V. S.; SINGH, S. K. Web crawler: a review. **International Journal of Computer Applications**, v. 63, n. 2, p. 31–36, 2013.

AL-KHULAIFI, M. **An Automated bibliographic network for the libraries of Riyadh, Saudi Arabia: a feasibility study**. 1997. Tese (Doutorado) – University College London. Disponível em: https://discovery.ucl.ac.uk/id/eprint/10104326/1/10017339.pdf.

KOSOKOFF, E. Tools and Trends: Internet Gopher and the Librarian. **Computer-Mediated Communication Magazine**, 1995. Disponível em: https://www.ibiblio.org/cmc/mag/1995/mar/kosokoff.html.

MARTELETO, R. M. Redes sociais, informação e conhecimento: inter-relações teóricas e práticas. **Informação & Sociedade: Estudos**, v. 17, n. 2, p. 1–17, 2007. Disponível em: https://ojs.uel.br/revistas/uel/index.php/informacao/article/view/1785.

MCMILLAN, G.; COX, J. Information at Your Fingertips: Organizing the Life Sciences Gopher. **Issues in Science and Technology Librarianship**, 1997. Disponível em: https://webdoc.gwdg.de/edoc/aw/ucsb/istl/97-spring/article2.html.

MILVUS RESEARCH. How do you integrate ranking signals in search engines?, 2023. Disponível em: https://milvus.io/ai-quick-reference/how-do-you-integrate-ranking-signals-in-search-engines.

MITRA, B.; CRASWELL, N. An Introduction to Neural Information Retrieval. **Foundations and Trends® in Information Retrieval**, Now Publishers, v. 13, n. 1, p. 1–126, 2018. DOI: 10.1561/1500000055.

MONIER, L. The AltaVista Web Search Engine. **;login:** v. 22, n. 2, 1997. Originally published April 1997. Disponível em: https://www.usenix.org/legacy/publications/library/proceedings/ana97/summaries/monier.html.

MORAES BRAZ, C. de; KATAGUE, G. P.; PINA JR., J. C. de. **PageRank**. São Paulo, SP, 2012.

MORESI, E. A. D. **Metodologia da Pesquisa**. Brasília, DF: Universidade Católica de Brasília, 2003. Disponível em: https://www.inf.ufes.br/~pdcosta/ensino/2010-2-metodologia-de-pesquisa/MetodologiaPesquisa-Moresi2003.pdf. Acesso em: 07 junho 2025.

NAAMHA, E. Q.; ABDULMUNIM, M. E. Web Page Ranking Based on Text Content and Link Information Using Data Mining Techniques. **ARO - The Scientific Journal of Koya University**, v. 12, n. 1, p. 29–40, 2024. DOI: 10.14500/aro.11397.

OKOLI, C.; SCHABRAM, K. A Guide to Conducting a Systematic Literature Review of Information Systems Research. **SSRN Electronic Journal**, 2010. DOI: 10.2139/ssrn.1954824. Disponível em: https://doi.org/10.2139/ssrn.1954824.

OLIVEIRA, M. V. *et al.* Page rank: o funcionamento da ferramenta de busca do Google. **Caderno de Graduação - Ciências Exatas e Tecnológicas - UNIT - Sergipe**, v. 3, n. 3, p. 73–73, 2016.

OLIVEIRA SOUZA, F.; SANTOS, J.; ALVES, C. O uso das tecnologias com base nas redes sociais para a divulgação de estudos e trabalhos científicos. **Perspectivas Online: Humanas e Sociais Aplicadas**, v. 11, n. 30, p. 55–67, 2021. Disponível em:

https://ojs3.perspectivasonline.com.br/humanas_sociais_e_aplicadas/article/view/2447.

PAGE, L. *et al.* **The PageRank Citation Ranking: Bringing Order to the Web**. Stanford, CA, 1999.

PALHARINI, E. P.; ZAMBERLAN, A. **Web Crawler para portais da área do Direito**. Santa Maria, RS, Brasil, 2023. Trabalho de Conclusão de Curso (Sistemas de Informação) – Universidade Franciscana (UFN). Disponível em: https://tfgonline.lapinf.ufn.edu.br.

PASQUALE, F. The Black Box Society: The Secret Algorithms That Control Money and Information. **Harvard University Press**, 2015.

PENNINI, A.; HERMONT, B. Pesquisa nas redes sociais da internet à luz da perspectiva sistêmica. **Organicom**, v. 15, n. 2, p. 76–89, 2018. Disponível em: https://revistas.usp.br/organicom/article/view/139342.

PICALHO, A. C. **Das bibliotecas aos buscadores: testando técnicas avançadas para a recuperação da informação em pesquisas por documentos na web**. 2023. Dissertação (Mestrado em Engenharia e Gestão do Conhecimento) – Universidade Federal de Santa Catarina, Florianópolis. Disponível em: https://repositorio.ufsc.br/handle/123456789/247585.

RESEARCHGATE. **O uso dos motores de busca na Internet: como se configuram as pesquisas de conteúdo na Web para a produção de trabalhos educacionais**. s.d. Disponível em: https://www.researchgate.net/figure/Figura-1-Arquitetura-e-funcionamento-dos-motores-de-busca-Fonte_fig1_300117039.

ROBERTSON, S. E. Theories and models in information retrieval. **Journal of Documentation**, MCB UP Ltd, v. 33, n. 2, p. 126–148, 1977. ISSN 0022-0418. DOI: 10.1108/eb026639.

ROUSSEAU, R. Daily time series of common single word searches in AltaVista and NorthernLight. **International Journal of Scientometrics, Informetrics and Bibliometrics**, Oostende, Belgium, v. 2/3, n. 1, 1998. Paper 2.

RUSSELL, S.; NORVIG, P. **Artificial intelligence: a modern approach**. 3. ed. Upper Saddle River, NJ: Prentice Hall, 2009. Disponível em:

https://repo.darmajaya.ac.id/5272/1/Artificial%20Intelligence-A%20Modern%20Approach%20(3rd%20Edition)%20(%20PDFDrive%20).pdf.

SCHWABER, K.; BEEDLE, M. **Agile software development with Scrum**. Upper Saddle River, NJ: Prentice Hall, 2002. p. 158. (Series in agile software development). 1st edition. ISBN 0130676349.

SCHWARTZ, B. **How Google Search Uses AI: RankBrain, Neural Matching, BERT and MUM**. 2022. Search Engine Land. Acesso em: 4 nov. 2025. Disponível em: https://searchengineland.com/how-google-uses-artificial-intelligence-in-google-search-379746.

SEYMOUR, D. T.; FRANTSVOG, D.; KUMAR, S. History Of Search Engines. **International Journal of Management Information Systems (IJMIS)**, v. 15, 2011. DOI: 10.19030/ijmis.v15i4.5799.

SHINTAKU, M.; CARVALHO, M. C. Ferramenta para descoberta de recursos (FDR): uma proposta para organização e recuperação da informação na Internet. **Ciência da Informação**, IBICT, v. 24, n. 2, p. 157–167, 1995.

SILVERSTEIN, C. *et al.* Analysis of a very large web search engine query log. **ACM SIGIR Forum**, ACM New York, NY, USA, v. 33, n. 1, p. 6–12, 1999. DOI: 10.1145/331403.331405. Disponível em: https://doi.org/10.1145/331403.331405.

SINGH, S. BERT Algorithm used in Google Search. **Mathematical Statistician and Engineering Applications**, v. 70, p. 1641–1650, 2021. DOI: 10.17762/msea.v70i2.2454.

SOUZA, D.; NEVES, L.; MELLO, G. Redes sociotécnicas e o acesso à informação na era das plataformas digitais. **Revista Ibero-Americana de Ciência da Informação**, v. 17, n. 1, 2024. Disponível em: https://periodicos.unb.br/index.php/RICI/article/view/53825.

SOUZA, R. R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação**, Escola de Ciência da Informação da UFMG, v. 11, n. 2, p. 161–173, 2006. Disponível em: https://www.scielo.br/j/pci/a/7tt9ykG8xTGbWsyYnDhmghr/abstract/?lang=pt. Acesso em: 7 jun. 2025.

SOUZA, R. H. d.; DORNELES, C. F. Analisando a eficácia do modelo vetorial de busca na ordenação de questionários*. In:* COMPUTAÇÃO, S. B. de (ed.). **Anais do Simpósio Brasileiro de Sistemas de Informação (SBSI)**. Lavras, MG, Brasil: Sociedade Brasileira de Computação, 2017. v. 13, p. 563–570. Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/6088. Acesso em: 8 jun. 2025. DOI: https://doi.org/10.5753/sbsi.2017.6088.

SULLIVAN, D. **A Complete Guide to the Google RankBrain Algorithm**. 2018. Search Engine Journal. https://www.searchenginejournal.com/google-rankbrain/227829/. Acesso em: 4 nov. 2025.

SULLIVAN, D. **FAQ: All About The New Google "Hummingbird" Algorithm**. 2013. Search Engine Land. Disponível em: https://searchengineland.com/google-hummingbird-172816. Acesso em: 6 abr. 2026. Disponível em: https://searchengineland.com/google-hummingbird-172816.

TAVARES, T. *et al.* **Os Motores de Busca numa Perspectiva Cognitiva**. 2009. Disponível em: https://repositorium.uminho.pt/bitstream/1822/9856/1/challenges_09_motores.pdf. Universidade do Minho, Instituto de Educação e Psicologia.

VERONICA: Very Easy Rodent-Oriented Net-wide Index to Computerized Archives. University of Oklahoma. 1990. Disponível em: https://www.ou.edu/research/electron/internet/veronica.htm.

VARIAN, H. R. Artificial Intelligence, Economics, and Industrial Organization. **NBER Working Paper Series**, 2019.

VIVEKAVARDHAN, J. Search Engines for User Centric Information Retrieval and Scholarly Communication. **International Journal of Advanced Library and Information Science**, v. 3, p. 201–211, 12/2015. DOI: 10.23953/cloud.ijalis.249.

WAZLAWICK, R. S. **Metodologia de Pesquisa para Ciência da Computação**. Rio de Janeiro: Elsevier, 2009.

WEIGHTED PageRank Algorithm Search Engine Ranking Model for Web Pages. **Intelligent Automation & Soft Computing**, v. 36, n. 1, p. 183–192, 2023. DOI: 10.32604/iasc.2023.032125. Disponível em: https://www.techscience.com/iasc/v36n1/49995/html.

WIGGINS, R. Gopher. **The Public-Access Computer Systems Review**, University Libraries, University of Houston, 1993. Electronic journal. Disponível em: https://uh-ir.tdl.org/server/api/core/bitstreams/0bc46bb9-b4e1-45fd-be9a-f3f94161e27f/content.

YUN, J. M. The role of antitrust in promoting innovation in the digital economy: A case study of Google. **Journal of Competition Law & Economics**, v. 14, n. 2, p. 311–338, 2018.