



UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

LAERCIO PONTIN JUNIOR

**SISTEMATIZAÇÃO DE UM MECANISMO DE BUSCA ESPECIALIZADO EM
CONTEÚDOS DAS SÉRIES INICIAIS DO ENSINO FUNDAMENTAL: uma
abordagem baseada em Mineração de Texto**

Tucuruí
2021

LAERCIO PONTIN JUNIOR

**SISTEMATIZAÇÃO DE UM MECANISMO DE BUSCA ESPECIALIZADO EM
CONTEÚDOS DAS SÉRIES INICIAIS DO ENSINO FUNDAMENTAL: uma
abordagem baseada em Mineração de Texto**

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Heleno Fülber
Coorientador: Prof. Dr. Adonney Allan de Oliveira Veras

Tucuruí
2021

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a)
autor(a)

P816s Pontin Junior, Laercio.
SISTEMATIZAÇÃO DE UM MECANISMO DE BUSCA
ESPECIALIZADO EM CONTEÚDOS DAS SÉRIES INICIAIS
DO ENSINO FUNDAMENTAL : uma abordagem baseada
em Mineração de Texto / Laercio Pontin Junior. — 2021.
109 f. : il.

Orientador(a): Prof. Dr. Heleno Fülber
Coorientador(a): Prof. Dr. Adonney Allan de Oliveria
Veras
Dissertação (Mestrado) - Universidade Federal do Pará,
Núcleo de Desenvolvimento Amazônico em Engenharia,
Mestrado Profissional em Computação Aplicada, Tucuruí,
2021.

1. Mecanismo de Busca Especializado. 2. Mineração
de Texto. 3. Ensino Fundamental. 4. Base Nacional
Comum Curricular. I. Título.

CDD 004

LAERCIO PONTIN JUNIOR

**SISTEMATIZAÇÃO DE UM MECANISMO DE BUSCA ESPECIALIZADO EM
CONTEÚDOS DAS SÉRIES INICIAIS DO ENSINO FUNDAMENTAL: uma
abordagem baseada em Mineração de Texto**

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Heleno Fülber
Coorientador: Prof. Dr. Adonney Allan de Oliveira Veras

Aprovada em 08 de novembro de 2021.

BANCA EXAMINADORA:



Prof. Dr. Heleno Fülber
UFPA - Orientador



Prof. Dr. Adonney Allan de Oliveira Veras
UFPA - Coorientador



Prof. Dr. Bruno Merlin
UFPA - Membro Interno



Profa. Dra. Sandra Regina Regina
UFES – Membro Externo

Dedicatória

Agradeço a Deus que me deu forças para concluir este projeto de forma satisfatória. A minha família, que abriu mão em muitos momentos de minha companhia para permitir esta caminhada. A minha esposa, Thacyane, que conseguiu proporcionar um equilíbrio em momentos truçulentos. A meu filho, Théo, o meu maior motivador e onde buscava forças para seguir. Aos meus amigos, em especial ao Vitor, que compartilhou viagens para Tucuruí, momentos cômicos e complicados dentro do programa de mestrado. Aos professores, em especial Dr. Heleno e Dr. Allan, que sempre me deram direcionamentos e conforto para a conclusão desta jornada.

RESUMO

A tecnologia está inserida nos mais diversos ambientes da sociedade, porém, em alguns momentos seu uso pode ser melhorado, como no caso dos mecanismos de busca web. Ferramentas de pesquisa na web, quando utilizadas por crianças do ensino fundamental, podem acarretar um ambiente complexo e que desfavoreça a aprendizagem. Tal ocorrência ocorre devido a pouca familiaridade de crianças em formular perguntas relacionadas ao que precisa aprender e a dificuldade em interpretar os resultados. Devido a essa dificuldade, este trabalho propõem um mecanismo de busca que aperfeiçoa a *string* de busca considerando o que rege a Base Nacional Comum Curricular. Para implementar este mecanismo é utilizada a mineração de texto, de forma a tornar o ambiente aprimorado para aprendizagem infantil. Como forma de avaliação foram utilizados alunos do ensino fundamental juntamente com a colaboração de professores do mesmo perfil. Deste modo é possível avaliar a utilização da ferramenta e concluir a ocorrência de melhoramento nos resultados de buscas educacionais ao se utilizar uma ferramenta que aplique técnicas de mineração de texto em um motor de busca.

Palavras-chave – Mecanismo de Busca Especializado; Mineração de Texto; Ensino Fundamental; Base Nacional Comum Curricular.

ABSTRACT

The technology is inserted in the most diverse environments of society, however, at times its use can be improved, as in the case of web search engines. Search tools on the web, when used by elementary school children, can lead to a complex environment that hinders learning. Such an occurrence is due to the little familiarity of children in formulating questions related to what they need to learn and the difficulty in interpreting the results. Due to this difficulty, this work proposes a search engine that improves the search string considering what governs the Base Nacional Comum Curricular. To implement this mechanism, text mining is used, to improve the environment for children's learning. As a form of evaluation, elementary school students were used together with the collaboration of teachers with the same profile. Thus, it is possible to evaluate the use of the tool and conclude the occurrence of improvement in educational search results when using a tool that applies text mining techniques in a search engine.

Keywords – Specialized Search Engine; Text Mining; Elementary School; Common National Curriculum Base.

SUMÁRIO

1	INTRODUÇÃO	10
1.1	PROBLEMA.....	11
1.2	JUSTIFICATIVA.....	12
1.3	OBJETIVOS.....	13
1.3.1.	Geral.....	13
1.3.2.	Específicos	13
2	REVISÃO SISTEMÁTICA DA LITERATURA	14
2.1	INTRODUÇÃO	14
2.2	DEFINIÇÃO DA BUSCA.....	15
2.3	FONTES DE DADOS.....	16
2.4	SELEÇÃO DO ESTUDO.....	16
2.4.1.	Avaliar a qualidade dos estudos.....	17
2.5	SUMARIZAÇÃO E QUADRO GERAL DE PESQUISAS.....	17
2.6	ANÁLISE E DISCUSSÃO	20
2.6.1.	Q1 - Quais motores de busca atendem crianças?	20
2.6.2.	Q2 - Quais características são abordadas nos motores para o público-alvo?	22
2.6.4.	Considerações	25
3	FUNDAMENTAÇÃO TEÓRICA	26
3.1	TECNOLOGIA NA EDUCAÇÃO	26
3.2	MOTOR DE BUSCA	27
3.2.1.	Motor de Busca na Aprendizagem	29
3.2.2.	Motores de Busca na Aprendizagem Infantil.....	31
3.2.3.	Interfaces Digitais Infantis	33
3.2.4.	Base Nacional Comum Curricular	35
3.2.5.	Mineração de Texto	37
4	METODOLOGIA	42
4.1	HIPÓTESES.....	42
	Hipótese Nula	42
	Hipótese de Pesquisa	42

4.2	QUESTÃO DE PESQUISA	43
4.3	TIPO DE PESQUISA	43
4.3.1.	Variáveis	43
4.3.2.	Sujeitos	44
4.4	ETAPAS DO DESENVOLVIMENTO	44
4.4.1.	Modelo de Trabalho	44
4.4.2.	Desenvolvimento.....	47
4.4.3.	Aplicação	48
4.4.4.	Validação	50
5	DESENVOLVIMENTO	52
5.1	IMPLEMENTAÇÃO DA FERRAMENTA EDUBUSCA	52
5.1.1.	Pré-processamento da base nacional comum curricular.....	53
5.1.2.	Mineração de texto.....	56
5.1.3.	Seleção de Dados.....	56
5.1.3.1.	Preparação dos Dados	56
5.1.3.2.	Indexação e Normalização de Dados	58
5.1.3.2.1.	Banco de Dados.....	58
5.1.3.2.2.	Carregamento de Dados	60
5.1.3.3.	Mineração de Texto e Cálculo de Relevância.....	62
5.1.4.	Motor de busca.....	65
5.2	PLANO EXPERIMENTAL.....	67
5.2.1.	Experimento de teste e adaptação	67
5.2.2.	Experimento final	68
6	RESULTADOS E DISCUSSÕES.....	72
7	CONCLUSÃO E TRABALHOS FUTUROS	79
	REFERÊNCIAS BIBLIOGRÁFICAS.....	81
	APÊNDICE A - QUESTIONÁRIO FINAL OBJETIVADO.....	87
	APÊNDICE B – ARTIGO PÚBLICADO XVIII ESUD	89
	APÊNDICE C – ARTIGO PÚBLICADO XVIII WEIBASE.....	101

1 INTRODUÇÃO

A tecnologia se tornou essencial em qualquer área de atuação. Ela trouxe ganho para os mais diversos segmentos, seja como meio, seja como fim. Nesse contexto, um divisor de águas na propagação das tecnologias digitais é a rede mundial de computadores. O número de usuários da internet passou dos 4,66 bilhões em 2021, assim pode-se considerar que ela é parte integrante da sociedade mundial (WE ARE SOCIAL; HOOTSUITE, 2021). Hoje, o maior fenômeno, do ponto de vista de todas as realizações da esfera da tecnologia, é a Internet (HUNÁKOVA; MIKULA, 2018).

A Internet virou uma autovia de veiculação da informação, assumiu uma plataforma onde os conteúdos são criados e compartilhados. É nítido o aumento da produção destes conteúdos nos últimos anos, mas como recuperar tais informações de forma relevante? Foi assim que surgiu a necessidade da criação de sistemas especializados na procura dessas informações nos servidores *web* espalhados pelo mundo, os denominamos motores de busca.

Um motor de pesquisa ou motor de busca (*Search Engine*) é um *software* que realiza a indexação de conteúdo disponíveis na internet, mais propriamente *websites*, permitindo que os seus usuários pesquisem qualquer tipo de informação através de palavras-chave, Costa (2011). Esta tecnologia é responsável por um grande avanço na busca de conteúdo na rede mundial de computadores, permitindo o consumo de informações, aprendizagem, entretenimento etc.

Ao pensarmos na educação, ampliou-se a acessibilidade das tecnologias nas escolas, podendo ser observada a implantação de laboratórios e salas de aula informatizadas. Tais investimentos prometem mudanças de paradigmas a fim de contribuir com a aprendizagem dos alunos. Porém, em muitos momentos o investimento na informática educacional é voltado para o *hardware* e se esquece da importância dos *softwares* focado ao aprimoramento e auxílio da educação.

Segundo Aresta, et al. (2008) a adesão à utilização de *softwares* educacionais pode ser indicadora do sucesso de uma nova abordagem do processo de aprendizagem, traduzida num novo modelo de trabalho, mais colaborativo e dinâmico, e no estabelecimento de relações interpessoais mais sólidas que ultrapassam o simples trabalho de grupo e se mantêm ao longo do tempo.

Tal adesão se torna ainda mais necessária ao observarmos eventos como a pandemia Covid-19, que acelera ainda mais a discussão de uma abordagem entre

aprendizagem aliada a tecnologia. Segundo o World Economic Forum (2020), com a Covid-19, 1,2 bilhões de crianças estão fora das salas de aulas e a solução para tal problemática é a utilização de aprendizagem de forma online.

Assim, aliando a facilidade do motor de busca em identificar informações, com a necessidade de novos *softwares* voltados para a educação surge a ideia de motores de busca voltados para o ensino. Tal ferramenta deve facilitar o encontro de informações voltada para a aprendizagem por um público que possui uma capacidade semântica ainda reduzida. Esse é o foco que se aborda ao longo desta pesquisa.

1.1 PROBLEMA

Hoje, a grande parcela dos sistemas de buscas não tem preocupação com o público infantil em seu retorno de busca. Por vezes trazem informações impróprias para a idade, bem como não possui filtro contra conteúdos adultos, propagandas e linguagens que não são habituais ao perfil do público. Sugestões de consulta podem exigir um nível de leitura avançada sobre temas complexos que as crianças têm dificuldade de compreender (PERA; NG, 2017).

Quando se fala em motor de busca, associa-se, por muito, o desenvolvido pela Google. Isso é demonstrado em pesquisa realizada na Alemanha, onde a Google tem participação de 86,59% do mercado de motores de busca, enquanto a ferramentas Bing tem 9,20% e Yahoo possui 2,63% (BABAYIGIT; GROPE, 2018). Devido a esta popularidade, torna-se implícito que grande parcela de alunos, que buscam aprendizado por meio de motor de pesquisa, utiliza tal plataforma como forma de apoio ao ensino.

Entre as ferramentas de busca na *web* existentes se ressalta que todas são de conexão aberta, o que gera o retorno de informações irrelevantes ou ainda negativas como a de conteúdos pornográficos ou provocativas. Assim, ao realizar buscas em ferramentas populares, crianças estão à mercê de toda e qualquer informação contida na *web* (Silva e Xexéo, 2013, apud Torres e Weber, 2011).

Azpiazu, Dragovic e Pera (2017), demonstram que as ferramentas existentes não trazem um ganho considerável para a aprendizagem. Afirmam ainda que ao realizar uma busca rápida nos motores existentes, depara-se com motores de *interface* confusa, que retornam pesquisas de publicidades, ou ainda funcionam com buscas limitadas. Para Verma, Minocha e Kochar (2014, pg. 22) “[...] a comunidade

social não tem um motor de pesquisa educacional dedicado que exhibe os resultados apenas a partir dos sites com conteúdo educacional.”.

Desta forma, este trabalho traz como problemática o uso de motores de busca no ensino de crianças. Isso se dá por uma ineficácia dessas ferramentas em seu retorno de busca quando voltado para aprendizagem não guiada ao público infantil. Tais características podem ser vistas nos retornos que geram, que por vezes apresentam conteúdos inapropriados para o perfil etário dos estudantes.

1.2 JUSTIFICATIVA

Existem inúmeros trabalhos propondo novas ferramentas de busca voltada para o público infantil, dentre os quais pode-se citar: Silva e Xexéo (2013) propõem uma ferramenta colaborativa, onde os próprios pais auxiliem seus filhos a identificar os resultados pertinentes. Verma, Minocha e Kochar (2014), propõem um motor baseado em multiagentes, um organismo vivo que se desenvolve a partir de cache do navegador, web logs e logs de pesquisa, traçando o perfil de busca do usuário.

A literatura aponta propostas diversas para motores focados na aprendizagem infantil, funcionando em um modelo assistido ou supervisionado. Porém, como fazer tal autonomia na busca, sem necessidade de terceiros na identificação do conteúdo buscado, de modo não apenas filtrar conteúdos inadequados, mas também adequar os resultados ao público-alvo?

Um dos modos de automatizar tal busca, e que é utilizado nesta pesquisa, é a utilização de um documento padrão. Para isso, o documento padrão existente, que rege os conteúdos e matérias, das mais diversas esferas da educação, é a Base Nacional Comum Curricular (BNCC), publicado em abril de 2017. Segundo o Ministério da Educação (2014, pg. 5), “[...] a BNCC foi preparada por especialistas de cada área do conhecimento, com a valiosa participação crítica e propositiva de profissionais de ensino e da sociedade civil.”.

Essa base é um documento de caráter normativo que define o conjunto orgânico e progressivo de aprendizagens essenciais que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica, de modo que tenham assegurados seus direitos de aprendizagem e desenvolvimento, em conformidade com o que preceitua o Plano Nacional de Educação (PNE).

Assim, este trabalho visa criar uma ferramenta de busca, baseada em mineração de dados e na BNCC, introduzindo assim agentes de busca focados em um documento padrão de conteúdo específico de aprendizagem de crianças. Com isso é possível gerar um ganho na aprendizagem do público-alvo, bem como qualidade nos conteúdos de buscas.

1.3 OBJETIVOS

1.3.1. Geral

O objetivo geral é desenvolver um mecanismo não supervisionado para customização de motores de busca na internet, baseado na Base Nacional Comum Curricular, que promova a adequação dos resultados da pesquisa ao público-alvo.

1.3.2. Específicos

Observado o objetivo geral do trabalho, identificou-se a necessidade de realizar os seguintes objetivos específicos:

- Elicitar requisitos e modelagem, de acordo com o perfil da educação básica;
- Determinar a técnica de mineração de dados mais adequada ao projeto;
- Desenvolver plataforma Web e motor de busca; e
- Validar a ferramenta e testes gerais.

2 REVISÃO SISTEMÁTICA DA LITERATURA

Este capítulo apresenta a revisão sistemática da literatura que tem como foco a identificação e avaliação de trabalhos relevantes ao trabalho proposto, bem como responder as questões da pesquisa de modo a nortear o objetivo do estudo.

2.1 INTRODUÇÃO

Faz-se necessário identificar e situar como as ferramentas de busca que atuam hoje no ambiente de aprendizagem de crianças, para isso a observação de pesquisas correlatas é importante. Esta revisão sistemática visa demonstrar o que existe, na literatura, referente ao comportamento dos motores de busca para crianças, quais as dificuldades e necessidades deste público. Com isso, pretende-se respaldar o objeto do projeto proposto, bem como validar sua importância, traçar o rumo a ser seguido durante a vida da pesquisa.

Azpiazu et al. (2017) afirma que ferramentas de pesquisa *web* não são necessariamente projetadas para facilitar a recuperação de recursos para esses usuários jovens, pois não auxiliam os usuários que possuem baixa alfabetização (dificuldade de compreensão e leitura). Vanderschantz e Hinze (2017) concluem que técnicas de construção de consulta podem ajudar retornando páginas e que uma diretriz de *design* pode melhorar a busca de informações deste público.

Assim, esta sessão tem por objetivo responder a seguinte questão de pesquisa (Q1): **“Como funciona um motor de busca que pretende atender a necessidade de aprendizagem de crianças?”**.

Para complementar a questão anterior, foram derivadas as seguintes questões:

- Q 1.1 – Quais motores de busca atendem crianças?
- Q 1.2 – Quais características são abordadas nos motores para o público-alvo?
- Q 1.3 – O que tem de falha nos motores de busca que atendem a população estudada?

Com as respostas destas questões, identificaram-se os motores de busca que são utilizados pelo público, assim como quais são as características que devem ser abordadas e quais as falhas podem acarretar o insucesso. Assim, torna-se possível traçar estratégias a serem seguidas no desenvolvimento da pesquisa, de modo a prever quais devem ser as características abordadas.

2.2 DEFINIÇÃO DA BUSCA

Com intuito de responder à questão da revisão sistemática as seguintes palavras chaves foram definidas, para a realização de pesquisa nas bibliotecas digitais:

- População:
 - Child;
 - Kid.
- Intervenção:
 - Search Engine.
- Aplicação:
 - Education;
 - Learning.

A *string* de busca definida para a busca inicial, a partir das palavras chaves e que trouxe um resultado considerável satisfatório foi:

("Child" OR "Kid") AND ("Search engine" OR "Search engine optimized") AND ("Education" OR "Learning")

Vale destacar que, além da definição e elaboração de *string* de busca, adota-se a técnica conhecida como *Snowball*, que é uma forma de amostra não probabilística utilizada em pesquisas sociais onde participantes iniciais indicam novos participantes (BALDIN; MUNHOZ, 2011). Essa mesma técnica pode ser abordada em trabalho escrito, onde através de citações e referencial, é possível identificar trabalhos correlacionados e com similaridade. Assim foi gerada uma interação entre trabalhos

correlacionados, o que possibilita a expansão do número de referencial considerado ao longo deste trabalho.

2.3 FONTES DE DADOS

Definiu-se como fonte de dados, livros, revistas, artigos, teses, dissertações, jornais e leis. A pesquisa considera a literatura publicada a partir do ano de 2013. Devem ser escritos em língua portuguesa e inglês. As buscas foram realizadas nos Periódicos da Capes e Scholar Google.

2.4 SELEÇÃO DO ESTUDO

Como a *string* de busca gerou uma quantidade expressiva de retorno que não eram pertinentes ao estudo, foram adotados critérios de inclusão e exclusão dos artigos, onde eram incluídos apenas:

- Trabalhos direcionados a otimização de motores de busca;
- Trabalhos que transmite a padrões de pesquisas realizadas por alunos na educação inicial;
- Trabalhos relacionados a mecanismo de busca otimizado para estudante;
- Trabalhos relevantes a buscas feitas por estudantes;
- Quando abordado características que identificam o uso de crianças em mecanismos de busca; e
- Mecanismos de busca focados a jovens e crianças.

Quanto ao critério de exclusão, foi observado:

- Trabalhos que não tenham métodos claros;
- Trabalhos cujo título difere do tema pretendido;
- Trabalhos que não tenha motor de busca ou tecnologia na educação como tema central;
- Trabalhos duplicados; e
- Trabalhos de motores de busca relacionados à área que não envolve a temática.

2.4.1. Avaliar a qualidade dos estudos

Para realizar um refinamento dos resultados, observou-se o título do trabalho. Como o retorno de trabalhos foi acima de 3.400 resultados, não sendo possível realizar o melhoramento de *string*, pois causou à perda de artigos importantes. Realizou-se então a fase de leitura de título como eliminação, reduziu de forma considerável os textos, resultando em 30 artigos relevantes.

Após análise do título, aplicou-se uma análise do resumo do texto e quando abordado a temática desejada, prosseguiu-se com o processo de leitura da íntegra. Deste modo, definiu-se 12 artigos importantes ao tema.

Como forma de complemento dos trabalhos, aplicou-se a técnica de *Snowball*, onde incorporou aos resultados 3 artigos, totalizando 15 artigos, abordando a temática desejada.

2.5 SUMARIZAÇÃO E QUADRO GERAL DE PESQUISAS

Como forma de elencar e demonstrar os trabalhos relevantes identificados, segue quadro discricionário das obras consideradas.

Tabela 1. Quadro de identificação de obras

Título do artigo	Autor(es)	Data de Publicação	Veículo de publicação
Child Search Framework: a collaborative information retrieval architecture to assist children in the search process	Sandra Regina Rocha Silva e Geraldo Bonorino Xexéo	2013	IEEE 17th International Conference on Computer Supported Cooperative Work in Design
A Multi-Agent Based Personalized Search Engine with Topical Crawling Capabilities	Verma, Disha, Minocha, Kanika e Kochar, Barjesh	2014	IUP Journal of Computer Sciences
PR Students' Perceptions and Readiness for Using Search	Mia Moody e Elizabeth Bates	2013	Journalism & Mass Communication Educator

Engine Optimization			
Model of the use of information technology in improving the ability of student (Experimental Study Tools Search Engine in SMA Negeri 1 Parung Bogor)	Roni Faslah	2016	Jurnal Ilmiah Econosains
Search Engine for Children: User-Centered Design	Tatiana Gossen, Michael Kotzyba e Andreas Nürnberger	2017	Datenbank Spektrum
The role of reading skills in the evaluation of online information gathered from search engine environments	Carolin Hahnel, Frank Goldhammer, Ulf Krohne e Johannes Naumann	2018	Computers in Human Behavior
Effects of student characteristics and question design on Internet search results usage in a Taiwanese classroom	Chuen-Tsai Sun, Shu-Hao Ye e Hsi-Chien Hsieh	2017	Computers & Education
BASE (Bielefeld Academic Search Engine)	Amelie Bäcker, Christian Pietsch, Friedrich Summann e Sebastian Wolf	2017	Datenbank-Spektrum
New query suggestion framework and algorithms: A case study for an educational search engine	I. Bahattin Vidinli e Rifat Ozcan	2016	Information Processing and Management
Online searching and learning:	Ion Madrazo Azpiazu,	2017	Information Retrieval Journal

YUM and other search tools for children and teachers	Nevena Dragovic, Maria Soledad Pera e Jerry Alan Fails		
Going online: Young children and teachers accessing knowledge through web interactions	Susan Danby	2013	Educating Young Children: Learning and Teaching in the Early Childhood Years
Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries	Dania Bilal e Meredith Boehm	2013	Qualitative and Quantitative Methods in Libraries
Using online data sources to make query suggestions for children	Maria Soledad Pera e Yiu-Kai Ng	2017	Web Intelligence
How Kids See Search: A Visual Analysis of Internet Search Engines	Nicholas Vanderschantz e Annika Hinze	2017	BCS Learning & Development
Design and Implementation of PIAK: A Personalized Internet Access System for Kids	Abrar Noor Akramin Kamarudim, Bali Ranaivo-Malançon e Nadianatra Musa	2017	Journal of Telecommunication, Electronic and Computer Engineering (JTEC)

Fonte: Próprio autor

2.6 ANÁLISE E DISCUSSÃO

Após concluir as etapas de planejamento, execução e sumarização dos trabalhos, buscou-se a leitura dos trabalhos selecionados de forma exploratória, a fim de discutir sobre as questões definidas na fase de planejamento. A principal dificuldade encontrada na revisão foi a quantidade de trabalhos com algum tipo de envolvimento com motores de busca, isso gerou uma grande massa de dados a ser analisada, o que necessitou de uma pré-avaliação através do título da obra.

Analisado os trabalhos e definido os de maiores relevâncias, torna-se possível discutir cada questão elaborada anteriormente.

2.6.1. Q1 - Quais motores de busca atendem crianças?

As análises dos trabalhos, desta questão, apontam que existem projetos direcionados a motores de busca com propósito infantil. No entanto, Silva e Xexéo (2013, pg.1) apontam que os atuais sistemas de colaboração remota foram criados para um público geral, de modo a atender a maior quantidade de pessoas. Com isso, acaba não focando na necessidade específica das crianças. Deixa ainda de observar as características do público infantil e suas peculiaridades.

As propostas de motores existem como é o caso do *Personalized Internet Access System for Kids* – PIAK, que foi desenvolvido por Kamarudim, Ranaivo-Malançon e Musa (2017). Essa proposta busca filtrar em duas etapas, a primeira elimina páginas indesejadas, por meio de lista negra e algoritmo. A segunda filtragem ocorre de forma colaborativa e baseada em conteúdo para recomendação, através de uma página principal segura (DMOZ).

O PIAK identifica o usuário através de um registro no sistema, e por meio desse mecanismo, verifica o conteúdo pertencente ao seu perfil (através da idade). Este sistema leva em conta ainda, o grau de escolaridade dos pais, para conseguir avaliar se o usuário poderá contar com auxílio dos pais durante a pesquisa. Ele considera ainda interface de interação entre o usuário e o sistema, adotando o tema da página do YouTube e a tecnologia *bootstrap*.

O modelo adotado por esta ferramenta aborda aspectos onde o usuário precisa estar com uma linguagem atendida pela DMOZ, pois ela é página principal, que será varrida pelo robô de busca. Além disso, o usuário precisa cadastrar o usuário,

para que ele consiga usar o sistema, tendo a necessidade de identificar a qualificação dos pais para auxiliar a busca (uma busca assistida).

Existe ainda estudo que desenvolve o motor através do modelo de pesquisa social da arquitetura integrada pais e filhos e visa melhorar o processo de pesquisa para as crianças. É o caso do *Child Search Framework* - CSF, elaborado por Silva e Xexéo (2013). Ele aborda a utilização de busca assistida, onde o pai supervisiona e decide quais os melhores resultados a serem apresentados para busca de seus filhos.

Esta proposta colaborativa é realizada por meio de remoto, para assim criar um aumento da cobertura do espaço de informação e diminuindo na redundância dos trabalhos. Aliado, é proposto ainda um agente inteligente, de modo a conseguir identificar padrões de escrita e definir o que está sendo buscado.

Verma, Minocha e Kochar (2014), propõem um motor de busca multiagente, que personalize os resultados com base no histórico de navegação do usuário e perfil explícito criado por ele. Em seu trabalho, apresentam a dificuldade de os motores de busca atuarem em um ambiente não controlado e direcionado a conteúdo educacional e a dificuldade de identificar o resultado desejado para este tipo de busca.

É proposto que as buscas iniciem em *sites* que possuam domínio específico da educação, e ainda que seja permitido ao usuário escolher qual tipo de local a busca ocorra (blog, site de notícia etc.). Foca ainda no retorno de informações baseados no histórico de atividades do usuário na Internet, utilizando esses dados para realizar o ranqueamento dos *sites* retornados. Por fim é destacada a necessidade de uma interface amigável (*viewfriendly*), preocupada com o dispositivo utilizado pelo usuário (*celular/desktop*).

Já Azpiazu et al. (2017), avaliam motores de busca existentes, voltados para a educação, como o YUM. Na identificação da busca de uma criança, aborda-se a necessidade adaptativa que o motor de busca deve ter para identificar possíveis erros de escrita ou de termos impróprios. Assim, o motor retorna ao usuário possíveis formas da *string* de pesquisa que interesse na busca.

Outra característica imprimida no YUM, que ganha destaque é a opção de consultas sugeridas, onde uma ferramenta *on-line* gera palavras-chave, dando alternativas de consultas para o usuário (espécie de auto completar). A legibilidade é

empregada ainda no trabalho, fazendo os retornos adequados de acordo com a capacidade de leitura das crianças.

Deste modo são notórios os rumos em que as pesquisas voltadas ao tema se encontram, bem como sua relevância. Infere-se que a grande maioria dos mecanismos necessitam de alguma forma, identificar o perfil de seu usuário. A preocupação com a parte da interface é ainda uma constante, pois sendo amigável, torna a interação com usuário algo mais agradável.

Características que adéquem a escrita e leitura das crianças são destaque. Porém, nenhum dos mecanismos busca um melhoramento do filtro, através de um documento base, que auxilie o mecanismo de busca a identificar o estudante deseja. Assim, esta proposta tornasse relevante, por não ter sido explorada e implementada até o momento.

2.6.2. Q2 - Quais características são abordadas nos motores para o público-alvo?

O estudo leva a uma série de características relevantes ao se desenvolver uma ferramenta de busca para o público. Porém, o que se observa é a necessidade de trabalhar de maneira a levar em conta a característica do público, bem como sua imaturidade. Sun, Ye e Hsieh (2017), em seu trabalho demonstra o modo que crianças formulam suas pesquisas, propondo até ensiná-los a formular de modo apropriado suas buscas.

Gossen, Kotzyba e Nürnberger (2017), que apontam ainda a necessidade de trabalhar, com crianças, em ambientes onde o usuário seja o centro, com um *design* fácil e divertido, proporcionando uma experiência agradável.

No trabalho é proposto um desenvolvimento que leve em conta a capacidade cognitiva, conhecimento e forneça apoio emocional para as crianças. A pontuar o requisito de suporte emocional, é apontado que crianças podem ficar frustradas rapidamente. Assim, é sugerido gerar sensações de sucesso sempre que possível e uma alternativa adequada para apontar possíveis falhas do usuário.

Leva-se em consideração o suporte de línguas, pois crianças, em muitos casos, ainda possuem dificuldade na escrita e leitura, o que pode gerar erros ortográficos. Além disso, pode haver dificuldade com o uso de teclado, outro fator de frustração da criança. Assim, é indicado que o mecanismo permita apontar correções

ortográficas e sugestão de consultas ao usuário. Pode, ainda, ter como complemento, um menu de navegação, que auxilia o público que ainda não possui facilidade em formular perguntas.

É abordado por fim o apoio de uma interface gráfica apoiada às habilidades das crianças, de modo que o uso seja fácil e intuitivo. O suporte ao julgamento de relevância dos sites recuperados é pontuado nesta literatura, devido a dificuldades de análise e identificação do público. Esse ponto vai ao encontro da necessidade da diversidade entre crianças, que podem ser influenciados pelas habilidades cognitivas, motoras, diferença de idades e indivíduos.

Ao observar o modo de se ranquear os motores de busca, Bilal e Boehm (2013), analisam o modo que se trabalham os motores de busca, e apontam ainda que existem melhorias a serem feitas. Nessa pesquisa, os autores apontam a necessidade um meio de classificação das consultas *on-line*, chegando a sugerir que sejam mediados por especialistas em informação, professores e/ou pais, de modo a julgar a relevância.

Vanderschantz e Hinze (2017), foca em identificar erros de buscas e quais abordagens são utilizadas nos motores, propondo melhorias em seu desenvolvimento. Entre elas está a utilização de técnicas de auxílio a construção de consultas. Assim, produz mais resultados, qualificados, e com efeito positivo por reduzir número de publicidade e informações indesejadas.

É apontada ainda a necessidade apoio as crianças por meio de diretrizes de *design*. Assim como em outros trabalhos, neste ponto se apresenta a necessidade de auxílio à reformulação a consulta. Indica a inda a possibilidade de inserir barras laterais que contêm informações relacionadas à consulta, que pode auxiliar os jovens pesquisadores.

2.6.3. Q3 – O que tem de falha nos motores de busca que atendem a população estudada?

Ao analisar as duas questões anteriores, junto com uma breve pesquisa nos motores especialistas para crianças, é notório falhas. Autores como Hahnel, Goldhammer, Krohne e Naumann (2018), identificam que nos motores de busca é exigida uma habilidade de leitura, para a seleção correta de resultados, pois não trazem por vezes os resultados desejados de forma direta.

Isso se dá pela necessidade de reconhecimento de palavras, compreensão semântica e compreensão da leitura, que gera a capacidade de verificar e descartar informações irrelevantes, bem como saber se informações adicionais de uma pesquisa devem ser levadas em consideração na consulta. Destacando que na educação, este déficit é ainda maior, sendo mais prejudicial.

Sun, Ye e Hsieh (2017), partem do princípio de que a formulação da pergunta é uma característica fundamental. Como comprovação treina o público infantil em como realizar buscas de forma mais precisa. Destaca ainda que o aluno que possui algum tipo de motivação, como um colega para realizar consultas em dupla, gera uma melhora nos resultados obtidos no mecanismo de busca.

Azpiazu, Dragovic, Pera e Fails (2017), analisa ferramentas focadas na busca de crianças e indica a necessidade de um motor onde alunos e professores interajam, direcionando resultados certos para a busca pretendida. Aponta como erro o retorno de publicidades e conteúdos indesejados e de forma incompreensível para o pesquisador. Levanta ainda o problema com as telas, que muitas vezes não adotam o perfil funcional em celulares e *tablets*, que são usados pelo público.

Semelhante a essa ideia, Silva e Xexéo (2013), propõem uma forma colaborativa, onde o público infantil tenha resultados de buscas filtrados pelos pais. Percebe-se assim, que as tarefas do motor de busca não estão otimizadas o suficiente, a ponto de autonomia para deliberar o que é um resultado relevante ou não.

Autores como Sun, Shu-Hao Ye e Hsieh (2017), trazem a ideia onde deve-se ter melhoramento no algoritmo de busca, levando em conta características: de interface, facilidade de compreensão. Apesar de existirem motores de busca preocupados com este item, acabam não cumprindo os demais requisitos para atender esta demanda.

Dania Bilal e Meredith Boehm (2013) propõem que os ranques existentes não levam a questão da idade com a relevância necessária, de modo a colocar o que importa e o que é esperado para os alunos, não causando assim a frustração por parte do pesquisador. Esta preocupação é de conhecimento geral, e alinhado com as sugestões anteriores, nos remete que crianças têm grande possibilidade de acessar resultados errados em sua pesquisa.

2.6.4. Considerações

Os trabalhos apresentam, de modo geral, a importância da qualificação no momento de realizar busca em um motor de pesquisa (Hahnel et al., 2018; Faslah, 2016; Sun; Ye; Hsieh, 2017). Aponta ainda que é possível realizar uma otimização no processo e que se pode treinar para isso (Faslah, 2016). Todos os cenários observados acontecem no meio acadêmico, em sua maioria com alunos secundários, que se aproxima com o objeto central da pesquisa aqui proposta. Destaca-se ainda o baixo índice de pesquisa envolvendo a temática de mecanismo de busca aplicado a educação básica, apesar de o acesso à internet por crianças ser crescente e funcionar como um mecanismo de auxílio na aprendizagem ativa.

Foram de grande importância os títulos escolhidos, pois tornou-se possível identificar alguns padrões entre os trabalhos, como a existência de deficiência no momento de busca de alunos na Web, a falta de um ambiente propício para tal (Azpiazu et al., 2017). Observou-se ainda que existem motores de busca que buscam traçar o perfil do usuário, ou ainda de ter uma busca colaborativa, de modo a atender o perfil das crianças (SILVA; XEXÉO, 2013; VERMA et al. 2014).

A temática abordada no projeto torna-se de grande relevância, e a pergunta de revisão é respondida, através de demonstrações e interesse no tema, demonstra que existem ambientes inovadores envolvendo a temática, porém que ainda não estão conseguindo gerar o impacto desejado. Inferem-se assim os seguintes aspectos mais relevantes a serem considerados na proposta deste trabalho, de acordo com a literatura aqui apresentada:

1. Filtrar conteúdos impróprios;
2. Possuir visual amigável;
3. Gerar recompensas satisfatórias;
4. Possibilitar o acesso através de várias plataformas (desktop, celular, etc.);
5. Classificar a busca através de especialistas da educação; e
6. Possuir sugestões de possíveis novas buscas.

Esta etapa da pesquisa foi publicada no Workshop de Educação e Informática Bahia-Alagoas-Sergipe (WEIBASE), evento pertencente a sociedade brasileira de computação e o artigo está disponível no apêndice C.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 TECNOLOGIA NA EDUCAÇÃO

É empírica o uso da internet dentro do processo de aprendizagem. Segundo Nair e Saras (2017), houve um desenvolvimento significativo na internet nos últimos anos, o que causou mudanças radicais na educação. Estas mudanças geram oportunidades na área, facilita assim a aprendizagem. Ganha-se maior notoriedade quando se observa a característica dos atuais alunos na fase escolar, denominados geração Alpha, pois ainda não experimentaram um mundo sem internet e desde muito jovens usam dispositivos inteligentes, o que altera sua forma de aprendizagem (NAGY; KÖLCSEY, 2017).

Devido a necessidade da nova geração, quanto a educação, influenciado pelo mundo ao qual vive, surgem diversas pesquisas inovadoras focadas na aprendizagem, por meio da tecnologia. Alguns trabalhos exploram o lúdico, pois resultam em aprendizagem cognitiva e motivacional (HUIZENGA et al., 2017), característica dos Alphas. Outros trabalhos buscam a exploração da educação híbrida, Azambuja, Cury e Grimoni (2016), apontam que o hibridismo é o resultado de métodos de ensino presencial e *on-line* com o foco em melhorar a experiência do aluno.

O método híbrido, por utilizar o método presencial de ensino em conjunto com ferramentas *on-line* para complementar o ensino, pode utilizar os motores de busca como ferramenta de auxílio. Este mecanismo está intrínseco na sociedade e tornou-se uma ferramenta ubíqua na *web*. Com uma ampla fonte de informações, eles fornecem acesso fácil e de qualquer finalidade, como educacional, ocupacional e privada (HAHNEL et al., 2017).

A ação natural da tecnologia dentro da educação pode ainda ser potencializada com fatores externos, como a Covid 19, o que antecipa e força uma abordagem da aprendizagem através da tecnologia. Na pandemia mais de 2 milhões de crianças e jovens ficaram em casa e a educação ajustou-se ao ensino à distância e de forma súbita a escola mudou-se para casa dos alunos e novas formas de ensinar foram necessárias. (CABRITO, 2021)

3.2 MOTOR DE BUSCA

Motor de busca, motor de pesquisa ou ferramenta de busca, é um sistema para procurar conteúdos alojados em *websites* ou documentos em bases de dados através de palavras-chave fornecidas pelo utilizador. Os motores de busca surgiram com a popularização da Internet, com a necessidade de um serviço que realize buscas da informação na rede, retornando resultados de forma rápida, eficiente e organizada. (SEEMS, 2020).

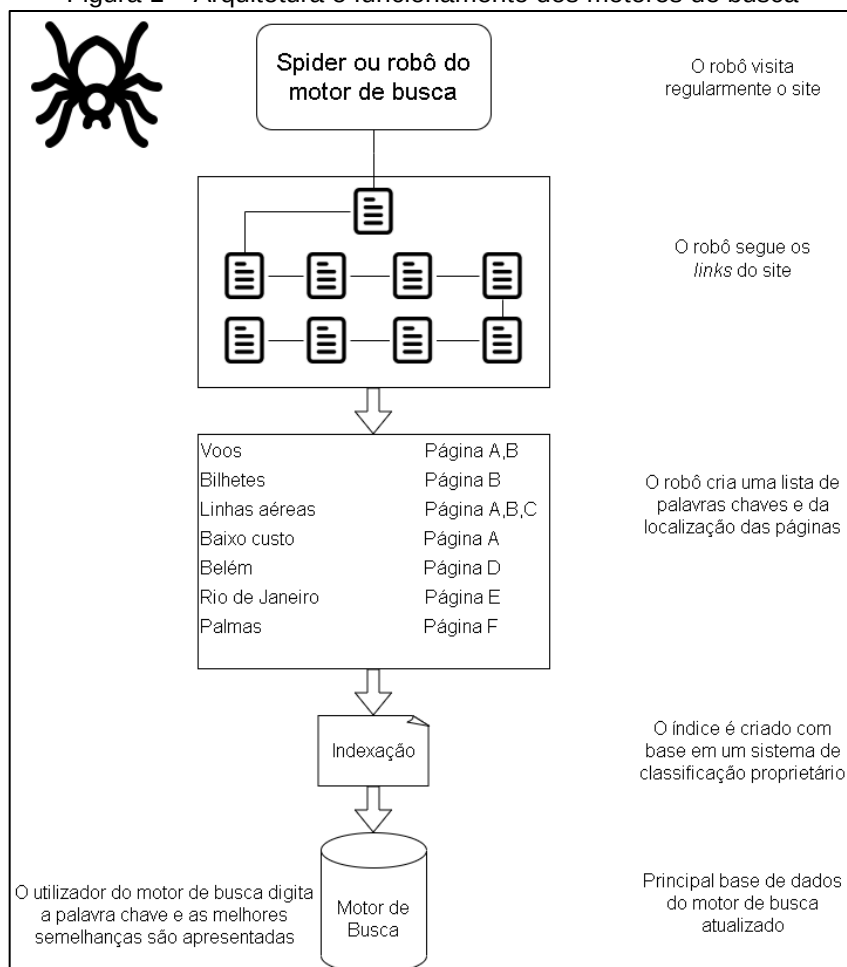
Logo empresas como a Google, o Yahoo, o Bing, o Lycos e o Cadê cresceram ao desenvolverem esse tipo de serviço. Para Rocha (2004):

Um motor de busca possui robôs, que visitam frequentemente as páginas da *web* e as entregam a um indexador de palavras, este indexador atualiza um índice, chamado de arquivo invertido, que é formado por três conjuntos: o conjunto das palavras indexadas, o conjunto dos endereços das páginas visitadas pelo robô e um conjunto composto por relacionamentos binários entre uma palavra do conjunto de palavras e um endereço do conjunto de endereços, em que cada relacionamento indica que a página do referido endereço contém a referida palavra. (ROCHA, 2004, p.110)

Segundo Correa, Warpechowski e Pinto (2014) “[...] a articulação entre as ferramentas tecnológicas e os conhecimentos escolares pode ser colocada em prática através de pesquisas escolares.”. Motores de busca agilizam pesquisas na rede mundial de computadores, através de algoritmos sofisticados, procurando resultados relevantes através de indexadores. Porém, cabe ao aluno realizar a pesquisa de maneira mais sucinta possível.

O mecanismo de busca funciona, de acordo com Correa, Warpechowski e Pinto (2014), rastreando links em páginas da internet e sendo classificado por índice, como apontado na Figura 1.

Figura 1 – Arquitetura e funcionamento dos motores de busca



Fonte: Adaptado de <http://evoluindonainformatica.wordpress.com/author/lanalbs>

Vale destacar que:

O acesso à Informática deve ser visto como um direito e, portanto, nas escolas públicas e particulares o estudante deve poder usufruir de uma educação que no momento atual inclua, no mínimo, uma “alfabetização tecnológica”. Tal alfabetização deve ser vista não como um curso de Informática, mas, sim, como um aprender a ler essa nova mídia. Correa, Warpechowski e Pinto (2014, apud Borba, 2001)

Assim, Correa, Warpechowski e Pinto (2014) observam que mecanismos de busca é um recurso importante para as pesquisas de conteúdo na internet, desde que utilizado com cuidado, em busca do tema desejado. Ao utilizar de forma correta, a pesquisa se torna ágil e objetiva, demandando menor esforço e tempo na ação de pesquisa. Com isso, uma ferramenta de busca, direcionada ao público em questão, gera um mecanismo ainda mais preciso e direcionado ao público fim.

Algumas estratégias de motores de busca, vem utilizando multiagentes, de modo a melhor qualificar seus resultados. Verma, Minocha e Kochar (2014) apresentam estes agentes como responsáveis por recuperar os resultados da pesquisa, monitorar as ações do usuário para priorizar os resultados, aprender de acordo com a escolha do usuário e sua atividade na internet. Após esta primeira etapa, o agente desenvolve uma técnica para aprendizagem, o que melhora o desempenho de recuperação do motor de busca através de ações prévias do usuário.

Para Chau et al. (2010) é crucial no motor de busca o indexador, responsável por construir índice de pesquisa para o repositório de páginas Web. De acordo com a frequência de palavras, da pesquisa, nas páginas “varridas” a lista de resultados de indexação é gerada. Assim, é exposto ao usuário, através de processamento prévio, uma classificação dos resultados das pesquisas.

3.2.1. Motor de Busca na Aprendizagem

Faslah (2016) avalia que a forma de busca, no Google, dos alunos de séries iniciais, tem por característica a dificuldade em construir termos a serem pesquisados nos motores. Os sites, em sua maioria, retornam dados que não devem ser acessados por crianças. Aponta ainda que a *web* é de suma importância para completar o ensino da sala de aula, por ser uma forma eficaz e barata de realizar atividades e complementar os estudos.

A maioria dos estudantes que usam mecanismos de busca consideram os retornos da pesquisa como as corretas, por muito sem questionar ou avaliar, o que entra em conflito com teorias da aprendizagem exploratória. Pesquisas longas e independentes exigem uma análise crítica dos resultados como parte de um processo que supostamente produz melhor resultado na aprendizagem. Para aumentar a motivação na exploração do conhecimento, muitos instrutores estão adicionando elementos de alto interesse aos alunos em seus materiais e atividades de ensino (SUN; YE; HSIEH, 2014).

Sun, Ye e Hsieh (2014), identificaram que os estudantes, mesmo quando em pares, não possuem aprendizagem ou habilidade para resolverem problemas, quando envolvem motores de busca. Por muito, os alunos não se aprofundam na pesquisa, para conseguir identificar o objetivo desejado. Em alguns casos, os alunos

buscam no tema algo que, associado com a relação interpessoal, pode ser discutido com os amigos.

Para que o aprendiz consiga obter conhecimento através de mecanismo de busca faz-se necessário uma certa habilidade no seu uso. Os motores de busca oferecem um período inicial de classificação, mas são os usuários quem definem se as informações listadas cumprem os requisitos da sua tarefa de pesquisa (HAHNEL et al., 2017). Assim, habilidade em leitura e interpretação na busca *web* é essencial para a satisfação nos resultados obtidos na pesquisa.

Hahnel et al. (2017), apontam que os efeitos globais da integração semântica e compreensão de leitura sugerem diferenças na largura e profundidade de processamento de informações. Os alunos que são capazes de capturar os contextos semânticos de textos fragmentados irão selecionar a informação apropriada. No entanto, alguns alunos só podem processar informações superficialmente até que encontre um que tenha grande semelhança semântica com objetivo buscado, enquanto outros comparam e relacionam cada *link* a partir de uma coleção limitada de modo a identificar a alternativa mais adequada.

Quando o cenário é direcionado para a geração Alpha, tem-se crianças em idade escolar primária, que estão em desenvolvimento, que se caracterizam como um palco onde estão aprendendo a raciocinar logicamente e tem dificuldades com o pensar abstrato. Nesta fase o seu entendimento é limitado ao concreto e conceitos físicos, ou seja, não devem ser usadas ideias abstratas em *interfaces* de buscas. Além disso, metáforas devem ser familiares para crianças e têm uma conexão com o mundo físico (GOSSEN; KOTZYBA; NÜRNBERGER, 2017).

Gossen, Kotzyba e Nürnberger (2017), apontam que os desafios das interfaces de pesquisa para crianças, podem vir com o suporte emocional, como apoio emocional e a sensação de sucesso. Quando o retorno retorna muito resultado, pode causar frustração, o que deve ser evitado através de mecanismos de ajuda. Outras características são o suporte à linguagem infantil (uma vez que estes usuários estão aprendendo a ler e escrever), ao cognitivo, relevância (crianças tem dificuldade em julgar relevância), interação simples e apoio a diversidade.

Logo, trabalhar com motores de busca voltados a crianças não é algo fácil, uma vez que diversos fatores devem ser observados. Não são apenas as habilidades das crianças que estão em formação, elas também podem ser diferentes das de outras crianças na mesma idade. Para enfrentar este desafio, um dispositivo de informação

formativo adaptável à idade é algo crucial. Infelizmente a *web* se adapta lentamente para as necessidades das crianças (SILVA; XEXÉO, 2013).

3.2.2. Motores de Busca na Aprendizagem Infantil

Devido a delicadeza do tema, bem como a sua importância, existem na literatura uma série de trabalhos que envolvem a temática. Vários mecanismos de buscas foram abordados, alguns observam a questão de *design*, como o apresentado por Gossen, Kotzyba e Nürnberger (2017), outros o papel das habilidades de leitura na avaliação de informações Hahnel et al. (2017).

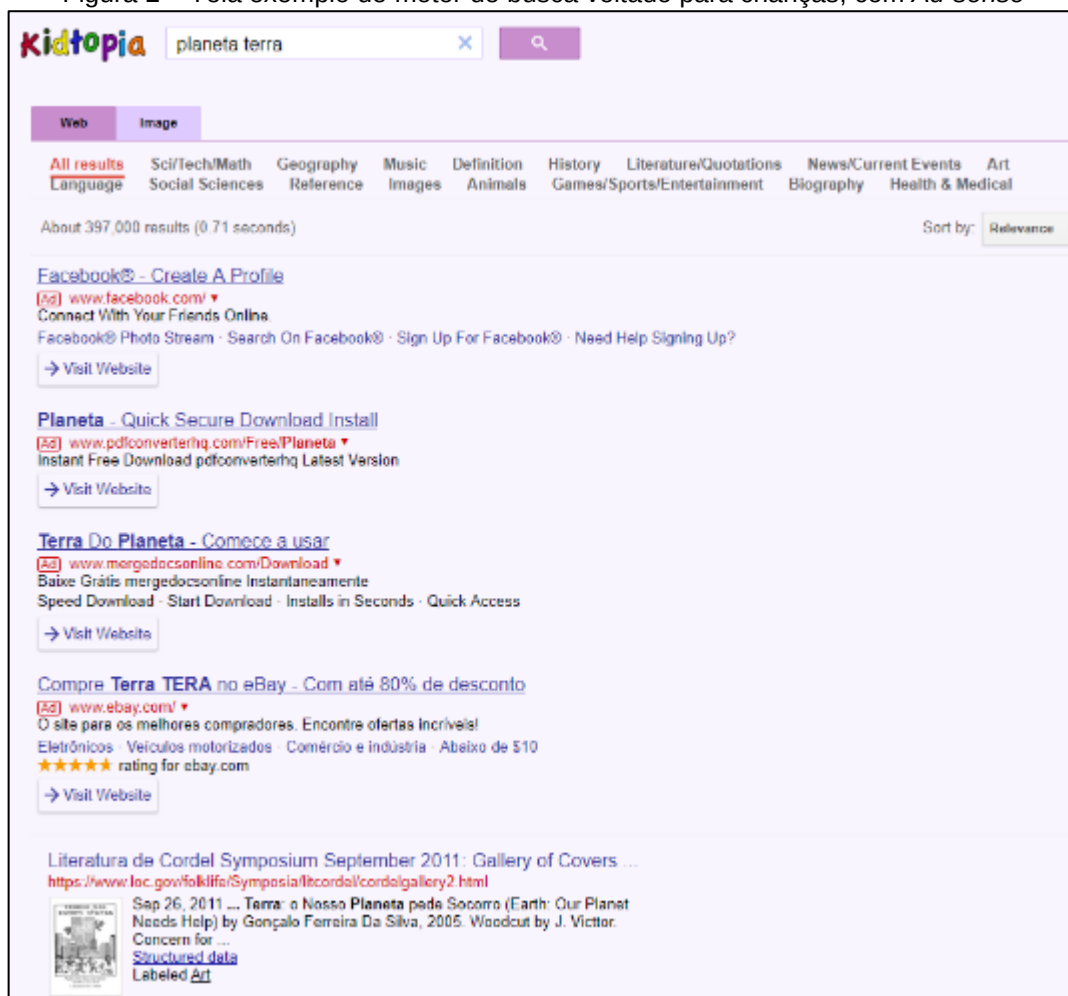
Rieh et al. (2016) demonstra que uma das principais dificuldades enfrentadas pelos alunos é a criação de consultas adequadas, que levarão as informações que de fato deseja. O fato de utilizarem a linguagem natural, ao invés de usarem palavras-chaves, que são esperadas pelos motores de busca, dificulta a identificação do resultado desejado.

Vanderschantz e Hinze (2017) indica que algumas técnicas devem ser adotadas para serem visualmente mais vantajosas para as crianças. Apontam ainda que a busca de informação de apoio das crianças pode ser melhorada por meio de diretrizes de *design*.

Alguns projetos já propõem mecanismo de busca para crianças, como: Kidtopia, Kidssearch, Ask Kids, entre outros, porém não atendem de fato a necessidade do estudante, como apresentado no trabalho de Azpiazu et al. (2017). Azpiazu et al. (2017, pg. 533) afirma que a “Dificuldade para recuperar recursos adequados / apropriados[...]”. Apontam ainda que, para 21% das consultas, as ferramentas de pesquisa orientadas para crianças, não recuperavam nenhum resultado ou os resultados recuperados não correspondiam ao que a criança esperaria.

Outro fato notório é que, de modo geral, as ferramentas desenvolvidas para pesquisa de crianças, são em sua essência em inglês. O que dificulta o acesso e aliado com a complexidade gráfica pode gerar um desinteresse de crianças. Apresentam ainda *Ad-sense*, como resultados primários das buscas, conforme Figura 2.

Figura 2 – Tela exemplo de motor de busca voltado para crianças, com Ad-sense



Fonte: Próprio autor, 2019.

Grande parte dos motores de busca possuem viés comercial, sendo por muito sustentados através de anúncios, os *Ad-blocks*, porém, esse elemento associado a busca de estudantes de séries iniciais tende a dificultar ainda mais um resultado satisfatório de pesquisa. Vanderschantz e Hinze (2017, pg. 3) afirmam que “[...] sites muito cheios de anúncios dificultam a encontrar informações. A inclusão de links de publicidade, é, portanto, provável prejudicial à pesquisa de informação bem-sucedida por crianças[...]”.

As características especiais das crianças são um desafio e deve ser considerado no desenvolvimento de motores de busca, incluindo o *design* de interfaces de pesquisa na *web* do usuário (GOSSEN; KOTZYBA; NÜRNBERGER, 2017). Interfaces gráficas devem ser bem desenvolvidas, amigáveis para crianças. Não é algo simples criar uma interface de pesquisa para crianças que agregue valor à pesquisa.

O ideal é fornecer aos usuários uma evolução interface de busca que se adapta às características individuais do usuário e permite mudanças. Não basta apenas flexibilizar as propriedades de elementos de interface do usuário, como cor e tamanho da fonte, mas também influencia elementos de interface do usuário e suas propriedades (GOSSEN; NITSCHKE; KOTZYBA, 2014).

Vale destacar ainda que a capacidade de avaliar e classificar a relevância de cada retorno de cada pesquisa compete ao motor de busca, o que é algo fundamental. Porém, segundo Vidinli e Ozcan (2016) a qualidade da consulta submetida é de extrema importância. Tem como finalidade criar sugestões e recomendações que ajudem o usuário a reformular sua busca, de modo a melhorar os resultados desejados.

Quando a pesquisa é realizada por crianças, a sugestão de busca, além de identificar melhoramento nas palavras chaves, deve considerar erros de ortografias, ou ainda excesso ou falta de letras em termos (GOSSEN; KOTZYBA; NÜRNBERGER, 2017). Como está em fase de desenvolvimento e aprimoramento da escrita, a capacidade de interpretação dos motores, deve ser capaz de compreender tais falhas.

Apesar da nova geração crescerem nativos no mundo digital e motores de busca, como Google, oferecerem assistência através da sugestão de palavras-chave e mecanismos de correção ortográfica, eles ainda podem ser um gargalo no momento da busca. A literatura apresenta os principais problemas enfrentados pelas crianças na recuperação da informação no modelo de pesquisa tradicional: ortografia, digitação, formulação de consulta (SILVA; XEXÉO, 2013).

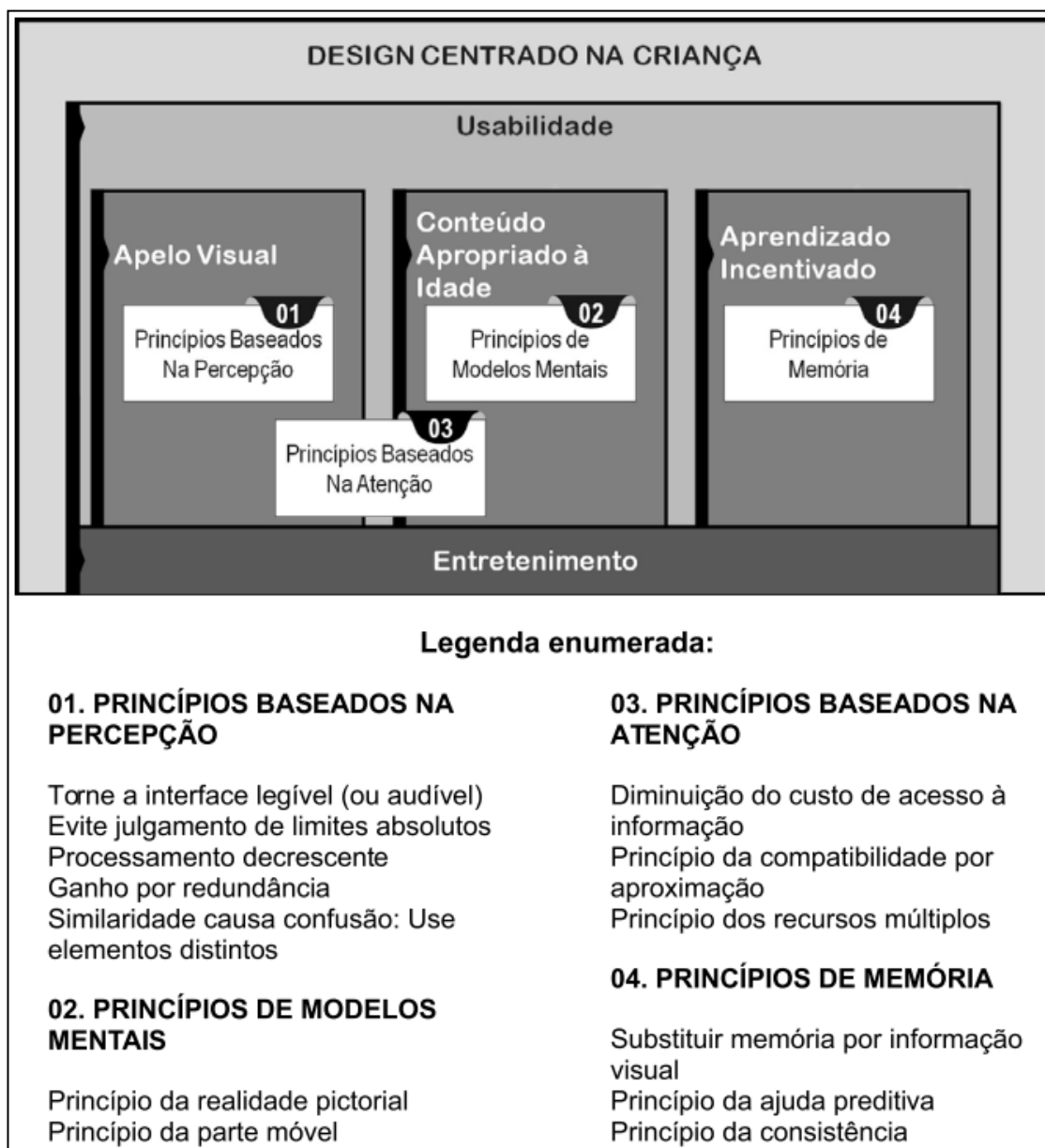
Silva e Xexéo (2013, pg.1) apontam ainda que os atuais sistemas de colaboração remota foram criados para um público geral, de modo a atender a maior quantidade de pessoas. Com isso, ele acaba não sendo focado na necessidade específica das crianças. Deixa ainda de levar em conta as características do público infantil e suas peculiaridades.

3.2.3. Interfaces Digitais Infantis

O *design* centrado no usuário, para Martins et. al. (2017), é uma das principais abordagens na contemporaneidade. Em seu trabalho, busca-se ainda, através de embasamento teórico, fundamentar o tangente a interfaces digitais para

crianças. Apontam, através da Figura 3, de forma agrupada, categorias com as características de usuário infantil.

Figura 3 – Características para usuário infantil



Fonte: Adaptado de Martins et. al. (2017)

Em cima das pontuações anteriores, é direcionado o alinhamento com foco em crianças. Para isso, deve-se atentar para a diferença de entendimento entre os diversos perfis dentre as crianças sendo algo lúdico, que atraia interesse. O apelo visual, através de um *design* de interface divertido, criativo, que cativa a criança e traga o interesse em permanecer no ambiente.

Apesar desses aspectos, não se pode deixar de considerar o critério da usabilidade, de modo que o perfil, ainda em desenvolvimento, seja atendido. Tais definições vão ao encontro do indicado por Gossen, Kotzyba e Nümberger (2017), que

apontam necessidade e suporte de memória ao trabalhar com crianças, bem como a necessidade de suporte à interação. Isso auxilia na satisfação do usuário e faz com que o número de interações continue crescendo.

Kamarudin, Ranaivo-Malaçon e Musa (2017), apontam ainda, que devido aos mais diversos meios de acessos atuais, a interface deve suportar multiplataforma. Dessa forma, se a busca for realizada por *desktop*, *smartphone*, ou qualquer outro tipo de dispositivo, não ocorrerá confusão em seu uso.

3.2.4. Base Nacional Comum Curricular

Ao longo desta pesquisa foi apontada, que para a ocorrência de pesquisa de informações por crianças através de motores de busca, é importante a classificação e identificação de conteúdo pertinentes. Para isso, a literatura sugeriu que pais e educadores avaliem e classifiquem estes documentos. Mas, como pilar desta pesquisa, é abordada uma característica autônoma de classificação, sugerindo-se a utilização da base nacional comum curricular para tal.

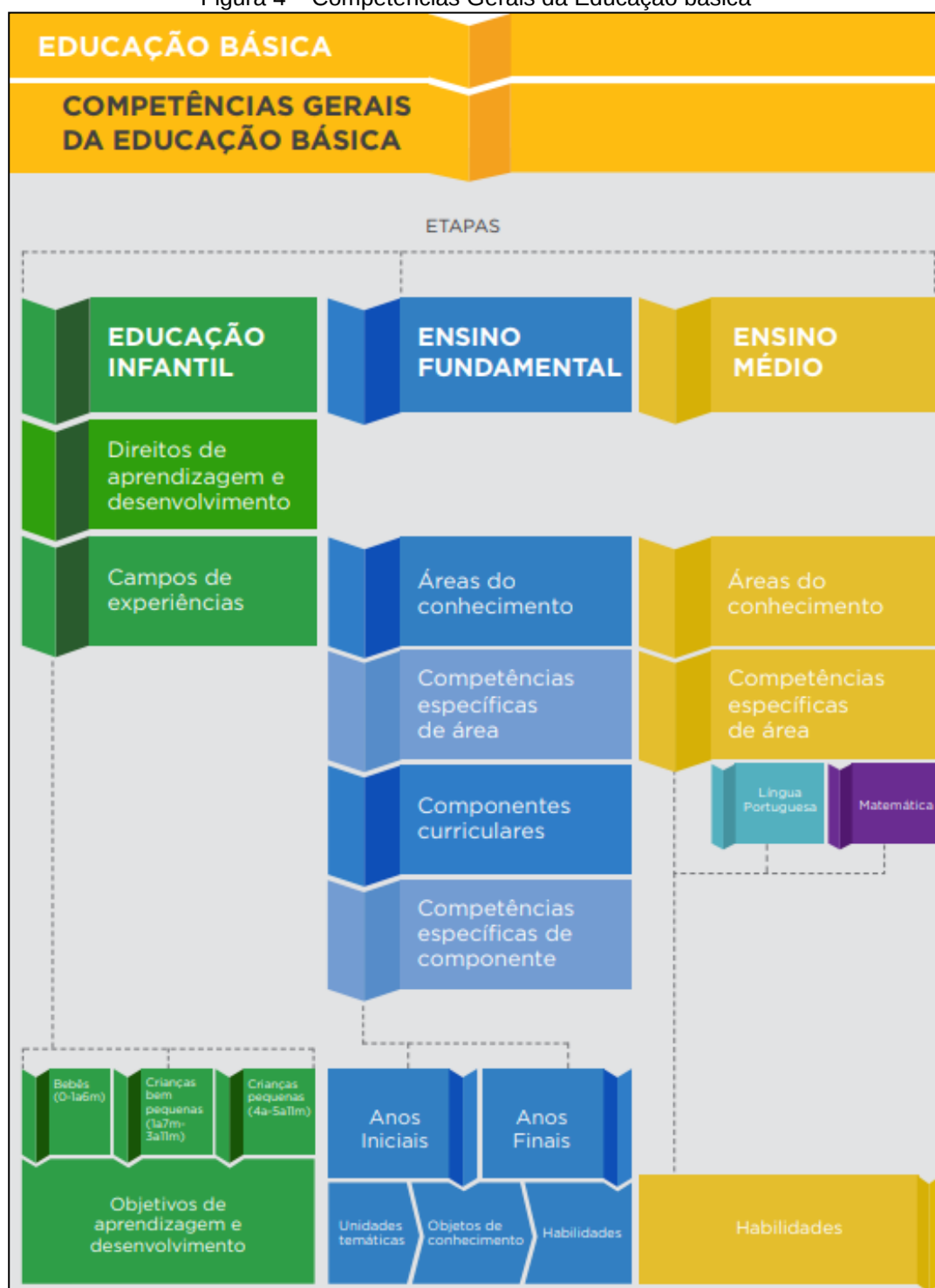
A Base Nacional Comum Curricular, segundo Ministério da Educação, é:

“[...] um documento de caráter normativo que define o conjunto orgânico e progressivo de aprendizagens essenciais que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica, de modo a que tenham assegurados seus direitos de aprendizagem e desenvolvimento, em conformidade com o que preceitua o Plano Nacional de Educação (PNE).”
(MEC, 2017, pg. 7)

Neste documento é apontado, pontualmente, o que os alunos necessitam aprender durante sua educação básica (direciona o que é ensinado para os alunos desta fase da educação). Nele são pontuadas, de maneira clara, as temáticas por matéria, os objetos de conhecimentos das mesmas e ainda as habilidades que o aluno terá com tal conhecimento.

Apresenta ainda a estruturação do ensino e como o mesmo deverá ser abordado, como apontado na Figura 4:

Figura 4 – Competências Gerais da Educação básica



Fonte: Ministério da Educação, 2017.

Estas características do documento poderão auxiliar e automatizar o processo até então executado por pessoas e que foi objeto de trabalhos estudados por esta pesquisa. Alinhada com as demais técnicas abordadas, será o ganho tecnológico deste trabalho, diferenciando dos demais aqui contido.

3.2.5. Mineração de Texto

A grande quantidade de informações disponíveis para serem buscadas na rede mundial de computadores exige *softwares* capazes de processá-los de forma automática na busca de conhecimento. Assim, Moraes e Ambrósio (2007, p. 2), apontam que (*Knowledge Discovery* - KD), “é um processo de análise de dados ou informações, cujo principal objetivo é fazer com a que as pessoas possam adquirir novos conhecimentos a partir da manipulação de grandes quantidades de dados.”.

Segundo Carvalho (2018), dentro da busca de conhecimento, a descoberta pode ser realizada através de duas áreas: “Descoberta de Conhecimento em Dados Estruturados (*Knowledge Discovery in Databases* - KDD) e a Descoberta de Conhecimento em Dados não Estruturados (*Knowledge Discovery from Text* - KDT).”, conforme a figura Figura 5.



Fonte: Adaptado de Moraes e Ambrósio, 2007.

O termo “mineração de texto” refere-se a análise de dados e informações contidas em textos escritos em modo livre, ou como comumente chamamos, em textos escritos em linguagem natural, ou seja, sem campos estruturados (RUIZ, 2016). O KDT utiliza abordagens das áreas de Recuperação de Informação, Processamento de Linguagem Natural e Descoberta de Conhecimento em Base de Dados de campos não estruturados e em textos não padronizados.

Devido à falta de estrutura, as informações “não são acessíveis para serem utilizadas por computadores” (CARVALHO, 2018, pg. 51). Para que se consiga abstrair o conhecimento a partir de um texto, segundo Moraes e Ambrósio (2007), faz-se necessária as seguintes etapas: seleção de documentos, definição do tipo de abordagem dos dados (análise semântica ou estatística), preparação dos dados, indexação e normalização, cálculo de relevância dos termos, seleção dos termos e pós-processamento (análise de resultados).

Entre as etapas, identifica-se que a análise estatística visa identificar e qualificar a importância da palavra está ligado ao número de vezes que aparece no texto. Segundo Moraes e Ambrósio (2007, pg. 8), seu processo envolve aprendizado estatístico a partir de dados, que normalmente inclui as etapas de codificação dos dados, estimativa dos dados e modelos de representação de documentos.” Ao realizar análise semântica busca a sequência de termos no contexto em que estão escritos, identificando sua função.

Segundo (REZENDE; MARCACINI; MOURA, 2011) “um dos maiores desafios do processo de Mineração de Textos é a alta dimensionalidade dos dados.”. Se tratando de textos, a existência de informações redundantes e/ou desnecessárias prejudica o desempenho e a qualidade dos resultados. Faz-se necessária a aplicação de métodos para extração e integração, transformação, limpeza, seleção e redução de volume destes dados (MORAIS e AMBRÓSIO, 2007).

Na preparação dos dados é fundamental, nele está contido o processo de Tokenização, conhecida ainda como Atomização, responsável por dividir um fluxo contínuo de caracteres em unidades mínimas, chamados de Token (REHMAN *et al.*, 2013). Assim, para Xavier, Silva e Gomes (2015), “o *token* é gerado descartando-se os caracteres de “espaço” e pode corresponder a uma palavra, símbolo ou pontuação.”.

A seleção de termos tem o objetivo de obter um subconjunto conciso e representativo de termos da coleção textual. O passo inicial é a eliminação de *stopwords*, “que são os termos que nada acrescentam à representatividade da coleção ou que sozinhas nada significam, como artigos, pronomes e advérbios.” (REZENDE; MARCACINI; MOURA, 2011). A eliminação da lista de *stopwords*, conhecida como *stoplist*, gera eliminação significativa nos termos do texto, o que diminui ainda o custo computacional, o que realiza redução de dimensionalidade.

A etapa de normalização consiste na “redução do léxico que identifica e agrupa palavras que possuem relação entre elas.” (BRITO, 2017). Segundo Morais e Ambrósio (2017), as variações morfológicas das palavras “são eliminadas através da identificação do radical de uma palavra. Os prefixos e os sufixos são retirados e os radicais resultantes são adicionados ao índice.”. Tal técnica é intitulada de lematização ou *stemming* e é exemplificada na Tabela 2.

Tabela 2 – Demonstração de *stemming*.

ID	Frase Normalizada	<i>Stemming</i>
1	Ideia genial	Ide gen
2	belo dia dirigir	bel dia dirig
3	ganhando aposta	ganh apost
4	perdemos jogo seremos eliminados	perd jog se elimin
5	valor novo carro subiu	val novo carr sub
6	perdi novamente aposta	perd nov apost
7	chove perigoso dirigindo	chov perig dirig

Fonte: Adaptado de Brito, 2017.

Em um documento, nem sempre as palavras contidas possuem o mesmo grau de importância. Segundo Morais e Ambrósio (2007):

Os termos mais frequentemente utilizados (com exceção das *stopwords*) costumam ter significado mais importante, assim como as palavras constantes em títulos ou em outras estruturas, uma vez que provavelmente foram colocadas lá por serem consideradas relevantes ou descritivas para a ideia do documento.

Assim, o cálculo de relevância dos termos em um documento é de grande importância. Morais e Ambrósio (2007), afirmam ainda que “o cálculo de relevância de uma palavra em relação ao texto em que está inserido pode basear-se na frequência dela, na análise estrutural do documento ou na sua posição sintática de uma palavra.”.

Para isso a avaliação através de estatísticas simples, como a frequência de termo (*term frequency* - TF), frequência de documentos (*document frequency* - DF) e a frequência inversa de documentos (*inverse document frequency* - IDF), permite classificar os termos a serem avaliados.

Uma vez ocorrido pré-processamento textual, deve-se classificar os mesmos. Para isso existe o método de classificação, utilizado no KDD, que é o

processo de aprender um modelo que descreve diferentes classes de dados, sendo estas predefinidas (ELMASRI; NAVATHE, 2011).

Para tal, utiliza-se o conjunto de treinamento, da amostragem, para construir o modelo de mineração de dados. As técnicas de classificação podem ser supervisionadas e não-supervisionadas. Dentre os modelos utilizados nos métodos de recuperação de informação, temos alguns destaques (WIVES, 2002 e EBECKEN; LOPES; COSTA, 2003, apud JACOBS, 2016):

1. Modelo Booleano: manipula um conjunto de documentos com “*and*”, “*or*” e “*not*” (expressões booleanas). Por exemplo, o usuário indica previamente um conjunto de termos que o documento precisa conter (faz-se uma intersecção) para que esse seja apresentado ao usuário;

2. Modelo espaço-vetorial: cada documento possui um vetor de duplas (palavra, peso), e cada vetor possui todos os termos da coleção, que podem ou não estar no documento. A partir disso, atribui-se o peso (um grau de importância) que pode ser calculado de diferentes formas (normalmente baseia-se no número de ocorrências);

3. Modelo probabilístico: também conhecido como Modelo Bayesiano, pois usa uma base matemática estatística do Teorema de Bayes, onde se calcula a probabilidade de ocorrer A, sendo que a única evidência que se tem é B;

4. Modelo difuso (*Fuzzy*): Neste modelo, os documentos também são representados por vetores de palavras, onde cada uma possui um grau de relevância. Assim, é criado um universo onde todos os elementos estão presentes em todos os conjuntos. Cada elemento tem uma relevância em cada um dos conjuntos, podendo ser zero - sem relevância alguma - ou um valor de relevância alto, sendo este valor um indicativo de quão importante para o documento é cada palavra;

5. Modelo busca direta: usa-se uma busca de *strings* em um documento ou em um conjunto de documentos, e como resultado são mostrados ao usuário apenas os documentos que contém as *strings* buscadas;

6. Modelo *Clusters*: consiste em encontrar os documentos que tratem de conteúdos similares, contando a quantidade de palavras semelhantes nos documentos e indexando por tópico (ou *cluster*);

7. Modelo lógico: torna-se necessário modelar os documentos com lógica predicativa, para que a aplicação possa decidir melhor a relevância de cada

documento, incorporando semântica ao processo de recuperação. É pouco utilizado fora do ambiente acadêmico devido ao grande esforço no trabalho de criá-lo;

8. Modelo contextual ou conceitual: este modelo considera que cada documento possui um contexto, assumindo que o texto, assim como a consulta do usuário, foi desenvolvido sobre um assunto específico. Uma vez identificado o contexto, pode -se recuperar informações relevantes em nível do tema abordado no texto, e não somente de palavras isoladas, como é feito nos demais modelos citados.

4 METODOLOGIA

Neste capítulo a metodologia da pesquisa é apresentada. Aborda-se as hipóteses consideradas, bem como a questão de pesquisa derivada delas. Apresenta-se ainda a classificação da pesquisa, bem como as variáveis a serem observadas e sujeitos participantes da pesquisa. Por fim, é demonstrado quais as etapas observadas ao longo da pesquisa, de modo a conseguir atingir o objetivo deste trabalho.

4.1 HIPÓTESES

Segundo Gil (2002, pg. 31), após definido o problema solucionável da pesquisa, deve-se “[...] oferecer uma solução possível, mediante uma proposição, ou seja, uma expressão verbal suscetível de ser declarada verdadeira ou falsa. A essa proposição dá-se o nome de hipótese.”

Existem diversas classificações para hipóteses, dentre elas estão a hipóteses de pesquisa e as nulas. Sampieri, Collado e Lucio (2013), definem as hipóteses de pesquisa como “Proposições provisórias sobre a ou as possíveis relações entre duas ou mais variáveis.”. Já as nulas são apontadas como negação ou refuto da relação entre variáveis, considerando assim a hipótese uma verdade.

Como forma de definir a relação entre as variáveis da pesquisa e tornar possível atestar ou refutar a solução proposta para o problema de pesquisa, definiu-se as seguintes hipóteses:

Hipótese Nula

Os motores de busca convencionais, não atendem as pesquisas educacionais de crianças em séries iniciais. (H0)

Hipótese de Pesquisa

Se motores de busca convencionais não atende a necessidade da educação infantil, então um motor de busca especializado, através da BNCC, é adequado para buscas educacionais de crianças em séries iniciais. (H1)

4.2 QUESTÃO DE PESQUISA

Em vista de objetivar este trabalho, será aplicada a seguinte questão de pesquisa:

Se a *string* de busca for alinhada a BNCC, um motor de busca é capaz de retornar os resultados pertinentes para crianças de séries iniciais?

4.3 TIPO DE PESQUISA

Esta pesquisa é objetiva, debatendo o estudo do conhecimento, seu paradigma é epistemológico. Quanto ao tipo de pesquisa, possui natureza aplicada e com fins descritivos. Quanto aos procedimentos, tem característica experimental, por verificar relações entre as variáveis das hipóteses e tem como instrumento a manipulação de condicionais e observação de resultados obtidos.

O método experimental, segundo Gil (2002), é baseado na submissão dos objetos de estudo, de modo a influenciar as variáveis, de forma controlada e conhecida pelo pesquisador, de forma a observar os resultados produzidos pela variável. Como esta característica é o que fundamenta esta pesquisa, este é o motivo de tal classificação.

4.3.1. Variáveis

Autores, como Prodanov (2013), apontam que existem dois tipos de variáveis, que elas são “[...] as características observáveis do fenômeno a ser estudado e existem em todos os tipos de pesquisa.”. Assim surge a divisão entre variáveis independentes (que influenciam e afetam as demais variáveis) e dependentes (que são influenciadas e que se busca explicar). Assim, de acordo com a hipótese, define-se:

- Independentes: Motor de busca especializado;
- Dependentes: Qualidade no retorno de busca na educação;
Idade;
Escolaridade; e
Desempenho escolar.

4.3.2. Sujeitos

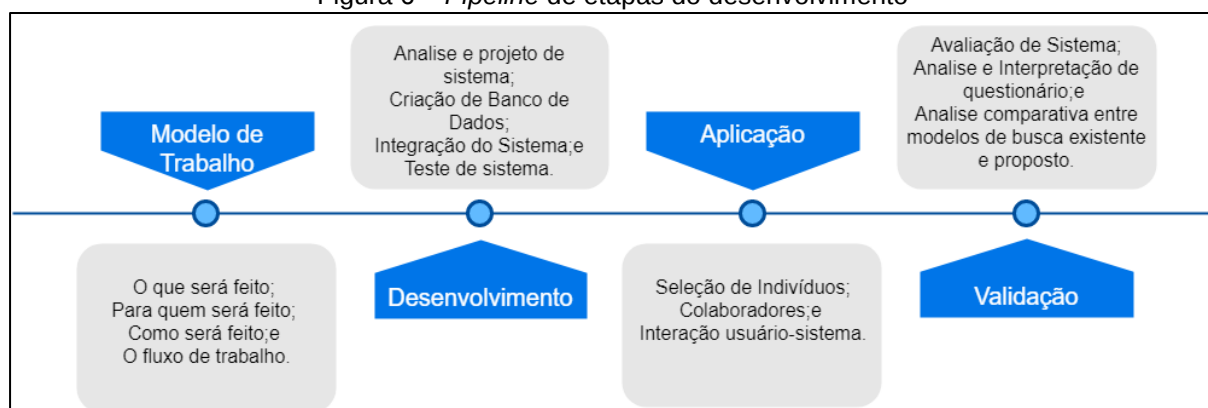
Devido à pesquisa ter sido desenvolvida em um período pandêmico (COVID-19), os sujeitos selecionados foram 15 alunos das séries iniciais, com idade entre 9 e 11 anos. Eles possuem acesso a um computador/celular com *internet* e possuem familiaridade com ela.

Houve também a participação de cinco professores colaboradores, com formação pedagógica e familiaridade na área (esses colaboradores são mais bem caracterizados no Item 5.3.1), de forma a auxiliar na aplicação da pesquisa e ainda a avaliar os retornos gerados nas buscas. São assim os responsáveis por avaliar os resultados, validando o modelo aqui proposto. A interação com os sujeitos ocorreu via *meet*, por ser a ferramenta utilizada na Escola Municipal Paulo Freire, Palmas -TO, local de aplicação deste trabalho.

4.4 ETAPAS DO DESENVOLVIMENTO

Para o desenvolvimento deste trabalho, foram delineadas 4 etapas a serem seguidas, conforme Figura 6. Tal sequenciamento tem papel de facilitar o entendimento e dimensionamento das atividades a serem realizadas durante a execução do projeto.

Figura 6 – Pipeline de etapas do desenvolvimento



Fonte: Próprio autor, 2019.

4.4.1. Modelo de Trabalho

Ao trabalhar com uma proposta de sistema computacional, a modelagem e definição dele possibilita estabelecer o que de fato será o produto. Assim, definir:

COMO, O QUE e PARA QUEM; ajudam a mensurar e identificar meios e papéis a serem abordados ao longo do projeto. Essa etapa consiste em abordar e atribuir tais valores ao mecanismo trabalhado neste projeto.

1. O que foi feito?

Foi desenvolvido um sistema computacional de adequação/auxílio a pesquisa na *Web* que retorna resultados de pesquisa mais apropriados para o público. Motores de busca funcionam com *strings* de busca (textos informados pelo usuário), que por muito não condizem com o que se pretende pesquisar. A melhoria no resultado ocorreu por meio de técnicas de mineração de dados e busca textual, que analisou a *string* de busca informada inicialmente e realizou, através de um documento padrão, uma adequação no texto base da pesquisa e na sequência fez a busca em uma ferramenta de busca consolidada na *Web*.

2. Para quem foi feito?

O público deste projeto são alunos do ensino básico, devido à pouca maturidade e expertise com a utilização de ferramentas de busca, além de estar desenvolvendo o pensamento cognitivo, o que dificulta na formulação de *string* de busca e a pouca capacidade de identificar os resultados relevantes para sua necessidade.

3. Como foi feito?

O mecanismo foi desenvolvido em ambiente *web*, adotando a responsividade para que o usuário o utilize independentemente do local de onde esteja acessando. A linguagem de programação para o *design* do *layout* da plataforma foram o HTML e o CSS, por darem suporte à ideia anterior. Adota-se o desenvolvimento *Web*, por ser mais fácil de acessar em computadores escolares, além de torna-se multiplataforma.

Sistemas *web* possuem a arquitetura cliente-servidor, onde no cliente é implementado processos que rodam no computador do usuário (e que foram definidas acima), já no servidor os processos ocorrem em locais remotos. Nesta pesquisa a linguagem de desenvolvimento para o servidor é o Python. Adota-se o mesmo pelo fato de já possuir ferramentas e bibliotecas que auxiliam na mineração de dados e buscas textuais, além de ser ágil para o desenvolvimento.

Dentre as bibliotecas do Python, utiliza-se:

- a. `urllib3` – Biblioteca que permite recuperar página *web*, de modo a possibilitar o trabalho futuro pela aplicação, neste trabalho será utilizado para varrer os dados contidos e identificar as expressões transcritas no site;
- b. `re` – Biblioteca que permite a realização do *splitter*, nesse processo são retirados caracteres que não são palavras;
- c. `nltk` – Biblioteca possui implementados métodos que auxiliam na retirada das *stopwords*, realização da *stemming*, entre outras funções;
- d. `pymysql` – Biblioteca responsável pela comunicação da aplicação com o banco de dados, permitindo a consulta e persistência do Python com o MySQL;
- e. `BeautifulSoup` – Biblioteca que identifica a página recuperada pela biblioteca `urllib3` (item a) de textual, o que possibilita a leitura e retirada de palavras, para processamento futuro; e
- f. `Urljoin` – Biblioteca que permite a manipulação do valor de url, de modo a permitir a busca em novas páginas *web*.

As técnicas de mineração de dados utilizam banco de dados para sua implementação. Assim, utiliza-se o MySQL, que é um sistema de gerenciamento de banco de dados relacionais de código aberto. Ressalta-se que tais linguagens são *open-source*, o que não gera custos para implementar o sistema computacional.

O fluxo de trabalho do mecanismo tem como início o usuário do sistema. Este sujeito, anteriormente definido como alunos do ensino fundamental, deve inserir no navegador, de forma textual ou por menu de navegação, o conteúdo que deseja buscar. Esta leitura da busca do usuário é processada por meio de mineração textual, que identifica os padrões de busca do usuário e a reformula de modo a adequá-la com a BNCC.

Uma vez o texto reformulado, o mecanismo realiza a busca no Google, de forma trazer da rede mundial de computadores os resultados. A busca ocorre no Google, pois ele é motor de busca mais difundido no mundo hoje. Possui ainda uma Interface de Programação de Aplicativos, em inglês, *Application Programming Interface* (API), que permite a integração de sites e aplicativos. Assim, a API retorna por meio de uma lista os resultados que podem reestruturados ser de acordo com a necessidade do desenvolvedor.

Uma vez a lista de retorno da busca ser gerada pela API, o sistema traz ao usuário os resultados oriundos da *string* adaptada pela ferramenta. Segue a Figura 7, onde demonstra o fluxo de funcionamento da ferramenta.

Figura 7 – Fluxograma do funcionamento do mecanismo proposto



Fonte: Próprio autor, 2019.

4.4.2. Desenvolvimento

Durante a etapa de desenvolvimento do *software*, adota-se a documentação através de diagrama de caso de uso e de entidade-relacionamento, padronizada pela Linguagem de Modelagem Unificada (UML). Através dele, permite-se identificar o fluxo do sistema, bem como o funcionamento do banco de dados e permite revisões do sistema por pessoas que não participem diretamente do projeto.

Criou-se ainda o banco de dados, suas tabelas e relacionamentos. Nesse momento a base de dados contém em sua estrutura um mapeamento de palavras identificadas, bem como a localização da mesma e o documento de onde foi retirado. Assim, torna-se possível identificar durante o processamento da busca o cálculo de relevância dos termos e ainda a seleção do mais apropriado para a pesquisa realizada.

O *design* da interface com o usuário, programação das funcionalidades do sistema, bem como sua integração com Google, bibliotecas de mineração e banco de dados também ocorrem neste momento. As tecnologias utilizadas são *open-source*, conforme abordado no Item 4.4.1.

A metodologia para desenvolvimento do projeto foi a XP, pois é focada em desenvolvimento rápido e com testes periódicos, de forma a ser rápida a adequação

de falhas. Ela passou por testes por parte dos professores e pedagogos dos alunos alvos desta pesquisa. Devido a seu conhecimento com o perfil dos alunos, estes colaboradores são capazes de indicarem falhas, erros ou ainda inconformidades do sistema. Assim, após validação e correção de todos os pontos indicados pelos especialistas, o sistema ficou apto para efetiva aplicação com usuário final.

Ao trabalhar com mineração de texto é necessário observar características apresentadas no Item 3.2.5. Nele é apontado a necessidade de seleção de documentos que guiaram a busca assistida do nosso modelo. Esta pesquisa utiliza como documento base a BNCC, como abordado ao longo do trabalho. Por buscar uma melhor adequação de resultados para pesquisa de crianças na *web*, ele tem como tipo de abordagem de dados a análise semântica.

A preparação ou pré-processamento dos dados, ocorre através dos termos contidos no documento base (BNCC), para tanto os conceitos de atomização é realizado por meio das bibliotecas apresentadas no Item 4.4.1 de forma a indexar os dados processados no banco de dados de acordo com apresentado anteriormente.

O pós-processamento (análise de resultados), é o que permite a otimização do resultado da busca do usuário do sistema proposto. Para tanto, o modelo de classificação é baseado no Cálculo de Relevância, apontado no Item 5.2.4. Definido o modelo, a aplicação final do usuário utiliza o mesmo, de modo a ser desenvolvido a parte gráfica do usuário e avaliado através da aplicação.

4.4.3. Aplicação

A etapa de aplicação da pesquisa é responsável demonstrar o modo que a pesquisa foi realizada, nesta etapa ocorre a interação com os sujeitos da pesquisa, identificados no tópico 4.3.2. Para tanto, o ambiente escolhido foi a Escola Municipal Paulo Freire em Palmas do Tocantins e seus professores e pedagogos (tidos como colaboradores). Os alunos foram selecionados pela equipe pedagógica da escola de aplicação.

Na sequência cada aluno indicou uma *string* de busca a ser utilizada no mecanismo proposto e no Google. Para tal, foram definidos 5 temas ao qual o aluno já estudou em sala de aula, definido pelo professor. Os temas definidos foram:

1. Ciclo das águas;
2. Reciclagem;

3. Trabalho e tecnologia;
4. Tipos/Gêneros de poemas; e
5. Poesia brasileira.

Devido a ocorrência da pandemia corona vírus – COVID-19, os alunos realizaram a indicação através do *meet*, sempre relacionado aos temas definidos pelo professor. É válido lembrar que se adotou esse modelo devido à dificuldade de trabalhar com pesquisa junto a alunos do ensino fundamental durante a pandemia. As escolas públicas, objeto da pesquisa, permaneceram por longos períodos fechados e quando retornaram as aulas foi via ambiente virtual de aprendizagem. Assim, o único modo de prosseguir com a pesquisa foi adotando uma metodologia de aplicação a distância.

Para avaliação do modelo, os sujeitos elaboraram cinco termos de busca, cada um com temáticas diferentes. Esses termos foram submetidos no motor de busca Google e no EduBusca (mecanismo proposto). Tal submissão foi feita para se tornar possível a comparação entre os resultados gerados pelo motor de busca mais utilizado na atualidade e a ferramenta proposta nesta pesquisa.

Após a submissão das *strings* de busca nos motores e com a geração da lista de resultados de cada um, foram selecionados os cinco primeiros *links* resultados do Google e os cinco primeiros *links* resultados do EduBusca, como forma de possibilitar a comparação e a avaliação entre ambas as ferramentas. Definiu-se a escolha dos cinco primeiros *links* resultados, de forma arbitrária, levando em conta que os primeiros resultados são os considerados mais relevantes pelo motor de busca.

Uma vez selecionados os *links* resultantes, eles foram direcionados aos professores, para que os avaliassem de forma cega. A forma cega consiste em ausentar a procedência do *link* para o professor e assim ele não ser capaz de identificar qual *string* gerou o resultado avaliado ou ainda se foi gerado pelo Google ou EduBusca. Desta forma é possível dar isonomia no processo de avaliação e validação da ferramenta.

4.4.4. Validação

Os dados para análise foram obtidos através de um questionário respondido pelo professor. Realizou-se uma análise quantitativa e qualitativa que considerou questões de facilidade linguística e de interpretação, se pertencem ao perfil do aluno, se está de acordo com a necessidade do mesmo e ainda se agrega valor a sua aprendizagem. O questionário pode ser observado no apêndice A, contendo quatro perguntas de múltipla escolha e que podem receber uma pontuação entre 3 e 0 pontos, a depender da escolha selecionada.

As questões abordam aspectos que possibilitam a adequação junto ao perfil do aluno. O questionário foi criado utilizando a ferramenta Formulário do Google, de modo a facilitar a interação do professor no momento de distanciamento social causado pela pandemia. Após avaliação dos professores é possível identificar a qualificação de cada *link* e assim validar se o trabalho possui um resultado positivo ou negativo.

A análise e interpretação da pesquisa foi realizada por meio da estatística do questionário, o que permite identificar padrões e constatações e qualificação dos resultados. Tanto nesta fase, quanto na coleta, foram levadas as características do público estudado, de forma a garantir a integridade da pesquisa. Outro fator que foi considerado na validação foram as dificuldades geradas pela COVID-19, como o distanciamento social. Esse fator não possibilitou a observação da utilização da ferramenta pelos alunos ou ainda uma análise de aspectos que necessitasse de contato direto com aluno e professor.

Concluído o levantamento de dados e a análise foi possível relacionar e alinhar o conteúdo tratado na pesquisa com os objetivos definidos de forma a criar maior relevância, através da ferramenta criada. Como forma de teste do modelo realizou-se uma primeira avaliação, com cenário reduzido de 3 alunos. Ela seguiu as etapas de aplicação com os alunos e validação dos resultados junto aos professores, através do comparativo entre os resultados gerados pelo Google e EduBusca.

No entanto, como foi uma etapa de avaliação inicial do funcionamento da ferramenta, nessa etapa ainda não existia o portal *web* implementado, sendo realizado o teste de funcionalidade direto pelo console da IDE que possibilitou, através da *string* do aluno, a sugestão da *string* otimizada, Figura 8. Esta avaliação inicial possibilitou

de modo geral a análise do modelo, das métricas e balanceamento utilizadas no cálculo de relevância e da adaptação da ferramenta para uma versão final.

Figura 8 – Demonstração dos resultados da primeira versão testada

```
In [16]: runfile('D:/Projetos-Python/consulta.py')
```

O que quer buscar? República do Brasil
A String ideal segundo BNCC é: história o nascimento da república no brasil e os processos históricos até a metade do século xx primeira república e suas características
contestações e dinâmicas da vida cultural no brasil entre 1900 e 1930

Fonte: Próprio Autor, 2020.

Na versão final do modelo, foram aplicadas as melhorias observadas no primeiro teste da aplicação, bem como implementada a parte gráfica do sistema. Nela a validação possibilitou a verificação do potencial total do sistema que foram mais bem detalhados nos tópicos posteriores, demonstrando como foi desenvolvida a pesquisa e quais os resultados obtidos ao longo do desenvolvimento do trabalho.

5 DESENVOLVIMENTO

Neste capítulo é apresentado o desenvolvimento da pesquisa. Nele é apontado a forma em que a ferramenta de busca (EduBusca) foi implementada. É nesse item ainda que se mostra o como foi realizada a execução da pesquisa.

5.1 IMPLEMENTAÇÃO DA FERRAMENTA EDUBUSCA

Para tal foram definidas as tecnologias, criadas as etapas pertencentes a mineração de texto, conforme Figura 9. A mineração aplicada a BNCC extraiu as informações que devem ser levadas em contas no momento de se realizar o melhoramento da busca no mecanismo de busca. É apresentado ainda o diagrama de caso de uso, bem como o sistema *web*, denominado EduBusca, que permite ao usuário ter a experiência com a ferramenta, tendo um melhor retorno na pesquisa *web*, de acordo com seu perfil.

Figura 9 – Fluxograma de trabalho de mineração de texto



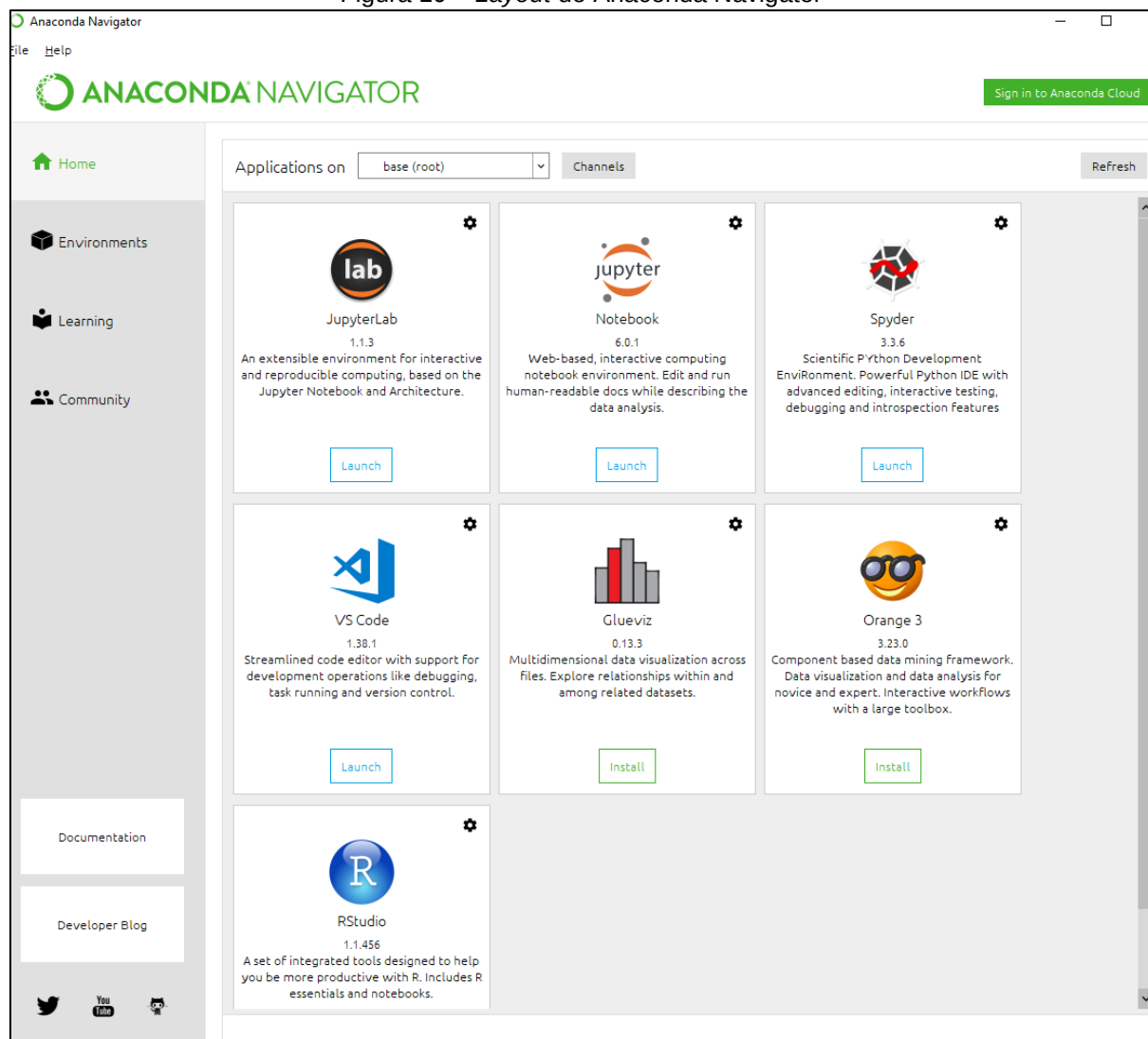
Fonte: Próprio Autor, 2019

Para implementação dos dispositivos desta pesquisa, é utilizada a linguagem Python através da interface gráfica de usuário (IDE) Anaconda Navigator. Tal IDE permite iniciar aplicativos e gerenciar facilmente pacotes, ambientes e canais conda sem a necessidade de usar comandos de linha de comando. Nesta ferramenta existem inúmeras ferramentas de desenvolvimento, como o Spyder.

O Spyder é um poderoso ambiente científico escrito em Python, para Python, e projetado por e para cientistas, engenheiros e analistas de dados. Ele apresenta uma combinação única da funcionalidade avançada de edição, análise, depuração e criação de perfil de uma ferramenta abrangente de desenvolvimento com a exploração de dados, execução interativa, inspeção profunda e belos recursos de visualização de um pacote científico.

Além disso, o Spyder oferece integração interna com muitos pacotes científicos populares, incluindo NLTK, RE, Pandas, IPython, URLLib3, SymPy e muito mais. Ele pode ser estendido ainda mais por plugins de terceiros, como uma biblioteca de extensão PyMySQL, permitindo que desenvolva suas interações com banco de dados MySQL.

Figura 10 – Layout do Anaconda Navigator



Fonte: Próprio Autor, 2019.

5.1.1. Pré-processamento da base nacional comum curricular

Para o funcionamento da técnica de processamento de dados, é fundamental um documento inicial de busca dos padrões a serem analisados, deste documento se abstrai a informação desejada e permite dar continuidade na abordagem estudada. Segundo Carvalho (2017), “tais documentos são recebidos e

automaticamente resumidos, internamente categorizados por um padrão de texto, e enviados automaticamente para os destinos adequados”.

Segundo Monteiro et al. (2006, apud Xavier, Silva e Gomes, 2015, p. 85) afirmam que ele pode ser dividido minimamente em três etapas: Pré-processamento; Análise e Extração do Conhecimento; e Pós-processamento. Uma vez pré-processado o documento é possível iniciar a mineração do texto. O documento da BNCC utilizado é uma tabela, formato XLSX, fornecido pelo site oficial do Ministério da Educação. Como está formatado para leitura e compreensão, sua estrutura possui uma série de informações que não agregam dados relevantes ao mecanismo de busca proposto (Cabeçalho, Rodapé, Informações Gerais do Documento, etc.).

O arquivo original possui 30 abas, dentre elas apenas 18 contêm informações referente a matérias escolares praticadas pelo ensino fundamental e cujo as colunas trazem dados pertinentes a cada tema da referida matéria e estão divididas em: Componente, Ano/Faixa, Unidades Temáticas, Objetivos de Conhecimento, Habilidades, Comentários e Possibilidades. Destas 18 abas, estão duplicadas as matérias, porém em uma das abas de mesma matéria possui duas colunas com material suplementar que auxiliam no entendimento do conteúdo exposto.

Por este motivo, optou-se por utilizar as abas que possui conteúdo suplementar, bem como os dados não suplementares. Desta forma, o arquivo original, foi reformulado, passando a não possuir mais dados de pouca relevância e abas com duplicidade, gerando um novo arquivo com 9 abas. Esta preparação, induz a redução da dimensionalidade dos dados a serem estudados e minerados.

Na sequência, foram retirados o caractere especial parêntese, para que não ocorra conflitos com o banco de dados no futuro. Por fim, as abas do documento foram renomeadas, retirando espaços e caracteres especiais, facilitando o trabalho durante a implementação da mineração de texto. As Figuras 10 e 11 apresentam uma comparação entre a planilha inicial e após o processamento.

Figura 11 – Demonstração de planilha não processada

	A	B	C	D	E
1					
2	Língua Portuguesa				
3	COMPONENTE	ANO/FAIXA	CAMPOS DE ATUAÇÃO	PRÁTICAS DE LINGUAGEM	OBJETOS DE CONHECIMENTO
4	Língua Portuguesa	1º	Todos os campos de atuação	Leitura/escuta (compartilhada e autônoma)	Protocolos de leitura

Fonte: Próprio Autor, 2019.

Figura 12 – Demonstração de planilha Pré-Processada

	A	B	C	D	E	F
1º		Língua Portuguesa	Todos os campos de atuação	Leitura/escuta - compartilhada e autônoma-	Protocolos de leitura	-EF01LP01- Reconhecer que textos são lidos e escritos da esquerda para a direita e de cima para baixo da página.
1						
1º		Língua Portuguesa	Todos os campos de atuação	Escrita -compartilhada e autônoma-	Correspondência fonema-grafema	-EF01LP02- Escrever, espontaneamente ou por ditado, palavras e frases de forma alfabética – usando letras/grafemas que representem fonemas.
2						

Fonte: Próprio Autor, 2019.

Além de facilitar a compreensão e implementação da mineração de texto, o pré-processamento do arquivo diminuiu o tamanho em *Kilobytes* (KB) do documento .xlsx. O documento original possui 3.188 KB, já o processado possui 645 KB. Com isso diminui o tempo e o uso de recursos computacionais necessários.

5.1.2. Mineração de texto

A mineração de texto é composta por uma sequência de atividades, de modo a permitir a consolidação dos resultados. Como já foi defendido na seção 3.2.5, este trabalho utiliza as etapas apresentadas na Figura 9, como forma de implementação da mineração de texto.

5.1.3. Seleção de Dados

O documento utilizado é uma planilha com diretrizes da BNCC, como forma de análise do arquivo é utilizada a técnica de varredura de linhas, colunas e abas. Para sua leitura utiliza-se a biblioteca WIN32COM.CLIENT que possibilita a abertura de arquivos no formato .xlsx, leitura de informações em suas diversas abas e manipulação de dados quando necessário.

Na sequência, como forma de recuperar os dados a serem processados nas rotinas de mineração de texto, utiliza-se a biblioteca *NumPy*, que possibilita a inserção das informações da planilha em *Arrays*, utilizados no processo de mineração de textos.

Uma vez feita a leitura da planilha, os dados são inseridos em estrutura de *Arrays* a serem processados em posterior pelas técnicas do KDT. Como este trabalho visa a identificação e qualificação de expressões linguísticas, ocorre durante a mineração uma análise semântica, para tanto, esses dados armazenados passarão por uma preparação (adequação).

5.1.3.1. Preparação dos Dados

Uma vez recuperados os dados a serem processados, faz-se necessário o seu preparo, que consiste na aplicação de técnicas que extraem e integram os dados, realizam sua limpeza e reduzem seu volume inicial. Esse desenvolvimento é útil para a não ocorrência de informações redundância e desnecessárias. Para tal, adotou-se as técnicas conhecidas como; Tokenização, retirada de *Stopwords* e *Stemming*.

O primeiro método aplica é o de Tokenização, com ele elimina-se espaços em branco entre as palavras de textos, bem como caracteres especiais avulsos como

(* , & , % , \$, # , @ , ! , entre outros). Com esta ação é reduzida a numerosidade de dados a serem trabalhados durante a mineração do texto.

Em Python, ao trabalhar com operações de expressão regular, existe a biblioteca RE, com vasto uso acadêmico. Este módulo fornece operações de correspondência de expressões regulares semelhantes. Entre suas funções, encontra-se a função *.compile*, que identifica espaço e caracteres especiais em expressões regulares. Essa função foi adotada na implementação de modo a reduzir a numerosidade textual durante o pré-processamento do texto.

Ao ser aplicado a quantidade de palavras do texto, que em seu formato original era de: 276045, com a retirada de espaços entre as palavras reduziu para: 239460 e com a retirada de todos os caracteres especiais houve a redução final para: 39786, o que corresponde a uma redução aproximada de 85,6% na numerosidade de expressões textuais a serem analisadas.

No segundo momento é realizada a retirada das ditas *Stopwords*. Para esta tarefa utiliza-se a biblioteca NLTK (*Natural Language Toolkit*). O NLTK “fornece interfaces fáceis de usar para mais de 50 corpora e recursos lexicais, juntamente com um conjunto de bibliotecas de processamento de texto para classificação, tokenização, *stemming*, análise...” (Documentação do NLTK, 2019).

Nessa biblioteca está implementado o método *.corpus.stopwords.words('portuguese')*, que contém uma lista prévia de palavras e termos considerados *Stopwords*. Com a utilização desta biblioteca, após a Tokenização do texto, a numerosidade de palavras para 25297 palavras, uma redução aproximada de 90,8% de palavras do texto e 37% sobre as palavras que passaram pela tokenização.

Por fim é implementado a técnica de *Stemming*, a qual gera uma redução na dimensionalidade de dados no banco de dados. Em sua implementação, faz-se presente a biblioteca NLTK, através do método *.stem.RSLPStemmer()*. Nele as palavras retornam ao seu sentido básico, indivisível e comum a todas suas derivações.

Abaixo é apresentado o quantitativo de palavras (Tabela 3) bem como a porcentagem da numerosidade inicial e pós-processamento do arquivo de mineração deste trabalho. O valor que inicialmente era de 276045 palavras teve uma redução de 90,8% do seu total, chegando à quantidade de 25297 palavras. Tamanha diminuição faz com que se obtenha uma melhor qualidade durante a execução do trabalho.

Tabela 3. Quadro de redução da numerosidade de palavras

	Quantidade	Porcentagem
Palavras Iniciais	276045	100%
Pós-Tokenização	39786	14,4%
Pós-Stopwords	25297	9,2%

Fonte: Próprio Autor, 2019.

Uma vez realizado o pré-processamento com êxito, é possível prosseguir com a atividade de indexação e normalização dos dados, conforme apresentado a seguir.

5.1.3.2. Indexação e Normalização de Dados

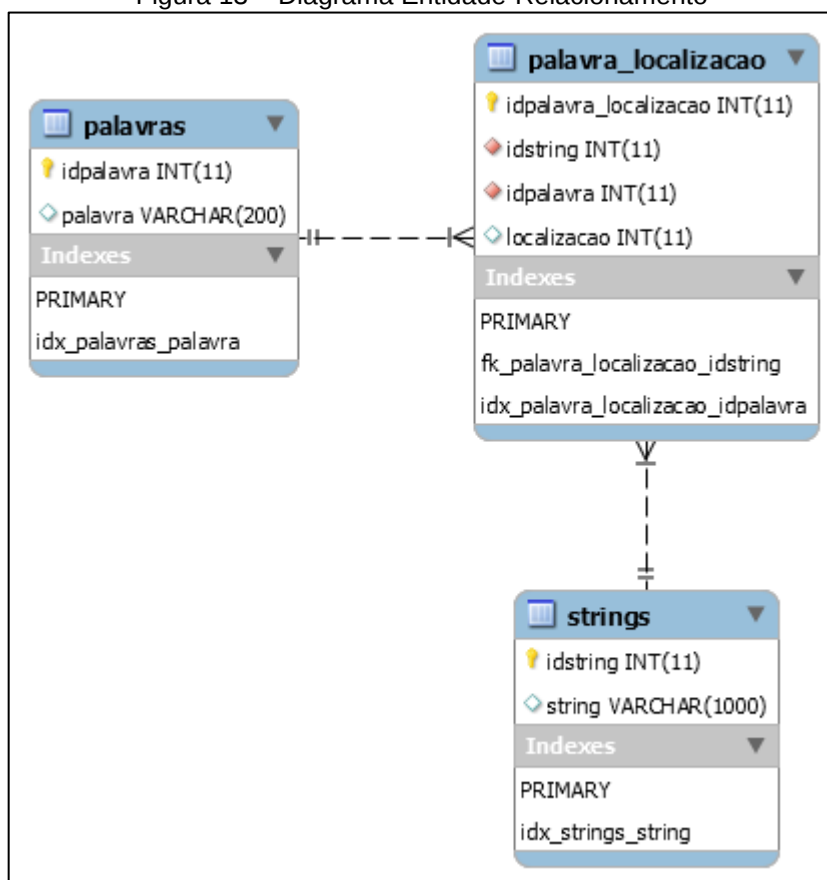
Após o pré-processamento do documento da BNCC, faz-se necessário a modelagem e criação da base de dados e ainda a sua inserção. Com estas etapas, é possível avaliar expressões textuais de forma estruturada, o que permitirá a execução da mineração de texto.

5.1.3.2.1. Banco de Dados

Para isso, foi utilizado o MySQL Workbench, que é uma ferramenta visual para o desenvolvimento de banco de dados, fornecendo modelagem de dados, desenvolvimento SQL, entre outras funções.

Para a pesquisa foi desenvolvido o diagrama de entidade relacionamento - DER, como forma de identificar os atributos, entidades e relacionamentos necessários para o trabalho. Na Figura 13 é apresentado o DER implementado no MySQL.

Figura 13 – Diagrama Entidade-Relacionamento



Fonte: Próprio Autor, 2019.

A entidade STRINGS é onde ficará armazenada a dita String Aperfeiçoada, ela será a *string* de retorno após a ocorrência da mineração de dados, que será aplicada ao Google, em busca da melhora nos resultados da pesquisa. É composto por: identificador: `idx_strings_string` (atributo: `idstring`) e a String Aperfeiçoada (atributo: `string`). Criou-se ainda o index (mecanismo responsável por tornar buscas, por um determinado atributo, mais rápidas) isso agilizará o retorno de consultas realizadas através do atributo `string`.

A tabela PALAVRAS é onde são inseridas as palavras pré-processadas na etapa de preparação de dados. Semelhante a entidade anterior, possui dois campos: identificador (atributo: `idpalavra`) e as palavras existente ao longo da BNCC (atributo: `palavra`). É verdadeira a afirmação de que palavras iguais não devem ser armazenadas, o que gera um banco de dados sem normalização, assim antes de inserir novas palavras é feita uma pesquisa na base, de modo a não duplicar informações. Devido a constante buscas pelo atributo `palavra` nesta tabela, também foi inserido um index: `idx_palavras_palavra`.

Como forma de vincular a String Aperfeiçoada com a parte textual que está a ela vinculada no arquivo da BNCC, adotou-se a tabela PALAVRA_LOCALIZACAO. Essa é a entidade responsável pela localização das palavras nos textos e a interligação entre as expressões textuais e a *string* que a representará ao final do processamento da mineração de texto. Possui o identificador (atributo: idpalavra_localizacao), os relacionamentos (atributos: idstring e idpalavra) e ainda o atributo localização, que indica a localização da palavra ao longo do texto.

Para melhor compreensão da localização, pode-se exemplificar uma *string* qualquer, como: “Pedro está jogando bola.”. Se dividir as palavras, e inserir em um *Array* X, ele ficaria da seguinte forma:

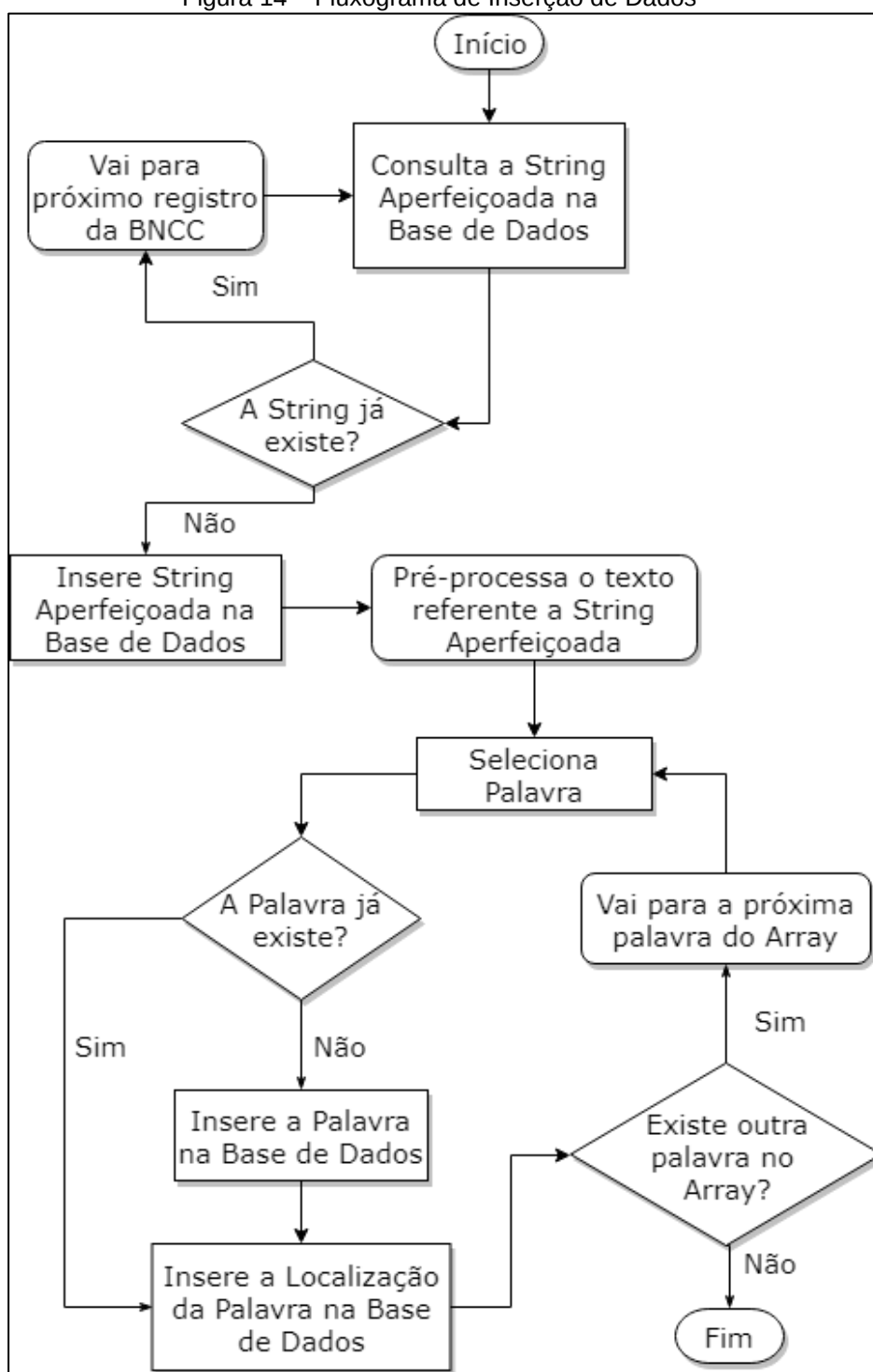
X = ['Pedro', 'está', 'jogando', 'bola']

Cada posição do *Array* está contida uma palavra (dado do tipo *string*) e desta forma é possível saber qual a localização de cada palavra na frase original. Este mesmo princípio é o adotado no atributo LOCALIZACAO, da entidade PALAVRA_LOCALIZACAO. Ele informa qual a posição original de uma determinada palavra em relação ao texto original.

5.1.3.2.2. Carregamento de Dados

Uma vez implementado o banco de dados, os dados foram inseridos através da aplicação em Python e nela utilizou-se a biblioteca *PyMySQL*. A regra de inserção segue o fluxograma da Figura 14.

Figura 14 – Fluxograma de Inserção de Dados



Fonte: Próprio Autor, 2019.

Com os dados armazenados na base de dados, torna-se possível realizar a mineração de texto e o cálculo de relevância dos termos, conforme veremos adiante.

5.1.3.3. Mineração de Texto e Cálculo de Relevância

Neste trabalho, como forma de mineração de linguagem natural foi adotado modelo de busca direta e dele gerou-se o cálculo de relevância de termos. Deste cálculo ocorre um balanceamento e os termos com maior pontuação são os considerados com retorno mais relevantes para o usuário.

Dentro do método de busca direta, os pesos foram calculados através de quatro métricas: a primeira avalia a frequência de vezes que a palavra é encontrada no texto, a segunda observa a localização do termo no texto, a terceira mede a distância entre os termos buscados dentro do texto e a quarta que qualifica o termo através do perfil informado pelo aluno. Como forma de balanceamento e normalização no cálculo de relevância é realizado o balanceamento dos valores, a depender da métrica utilizada.

Na métrica de frequência de palavras identifica a quantidade de vezes que cada palavra do termo informado de entrada na mineração é identificada nas expressões textuais, presentes no banco de dados (tabela: PALAVRA_LOCALIZACAO). Cada vez que se identifica a palavra analisada, ela recebe uma soma de mais 1 (um). Quando ela não é mais identificada é passada para a próxima palavra do termo e o processo anterior é repetido.

Ao finalizar a avaliação de todas as palavras do termo de entrada, informado pelo usuário é realizada a normalização dos valores, cujo a escala foi definida entre 0 e 1, arbitrariamente, onde o valor 1 é o mais relevante possível. A normalização ocorre através da fórmula abaixo:

$$X = Y/Z$$

Onde X é a nota normalizada (entre 0 e 1), Y é a quantidade de vezes que a palavra avaliada ocorre e Z é a quantidades de vezes que ocorreu a palavra com maior frequência.

Já a avaliação da localização do termo dentro do texto é a métrica utilizada para identificar a distância entre as palavras do termo de entrada do usuário e a sua localização no texto avaliado durante a mineração. Quanto mais próxima a palavra do início do texto, maior sua relevância e mais qualificado o termo para o usuário. Esta técnica utiliza-se do atributo LOCALIZACAO, contida na entidade PALAVRA_LOCALIZACAO.

À medida que a palavra fica mais distante do início do texto, soma-se mais 1 (um), o que faz com que os textos mais longes do início possuam valor maior. Como quanto maior o valor, pior ele é para o retorno, a normalização do dado deve considerar que o menor valor da localização significa que é melhor o resultado.

Assim, a normalização dos valores obedece a escala de 0 até 1, com o valor 1 sendo a melhor localização possível. Tomando por base o valor 1 sendo o melhor, definiu-se a fórmula de normalização abaixo:

$$W = T/K$$

Onde W é a nota normalizada (entre 0 e 1), T é a distância da localização da palavra em relação ao início e K é a menor distância de localização existente entre as palavras pesquisadas.

A terceira métrica aplicada mede a distância entre os termos, ela executa um cálculo que identifica a distância entre as palavras informadas pelos usuários e suas localizações no texto base da BNCC. Quanto menor a distância entre os termos, menor o valor da distância, o que aponta maior qualidade no resultado da busca. Este cálculo também é baseado no atributo LOCALIZACAO, da tabela PALAVRA_LOCALIZACAO, e usa a mesma normalização utilizada na métrica de localização.

$$H = J/L$$

Onde H é a nota normalizada (entre 0 e 1), J é a distância da palavra avaliada em relação a outra palavra pertencente a pesquisa e L é a menor distância entre as palavras pertencentes a pesquisa.

A quarta e última métrica adotada é a de verificação do perfil (série escolar) que o aluno pertence. Os termos armazenados no banco de dados estão qualificados por séries, de acordo com o indicado na BNCC. Ao iniciar a utilização da ferramenta, o aluno informa ao sistema qual série estuda, conforme a Figura 15. Desta forma a métrica busca qualificar os termos retornados através da aproximação do perfil informado na base comum e do informado pelo aluno.

Figura 15 – Identificação de perfil do aluno



Fonte: Próprio Autor, 2021.

Está métrica tem peso 0,5 e é menor que as demais, isso ocorre devido a casos em que alunos estejam estudando assuntos de perfis diferentes ao indicado pela BNCC. Por mais que essa diferença entre perfis ocorra, quanto mais próximo do perfil atual do aluno o termo necessário esteja, mais bem qualificado (mais apropriado) ele deve estar na apresentação do aluno.

A verificação de perfil consiste em subtrair o valor do perfil do aluno e o valor do perfil do sistema, a distância entre os perfis (a subtração entre ambos) é denominada P. Para equalizar o valor com seu valor máximo, utiliza-se a fórmula abaixo onde D é a nota normalizada (entre 0 e 0,5):

$$D = 0,5/P$$

Uma vez todas as métricas executadas, é gerado o peso de cada retorno da mineração. Quanto maior a nota da métrica, melhor sua acurácia. Assim cada consulta executada pelo usuário terá 4 métricas de avaliação. Todas as métricas possuem um peso durante a avaliação. Estes pesos foram definidos arbitrariamente após avaliação do comportamento do sistema e indicação de maior qualificação pelos professores colaboradores.

Somados os valores das métricas, gera-se o *score* que é vinculado ao termo do banco de dados a qual se examinou, dita String Aperfeiçoada. A String que possuir o maior valor de *score*, possui maior relevância e aproximação com o objeto de pesquisa do usuário. Assim, conclui-se a classificação e a etapa de melhoramento

do termo de busca para o aluno e inicia-se a pesquisa no motor de busca Google, por meio de sua API.

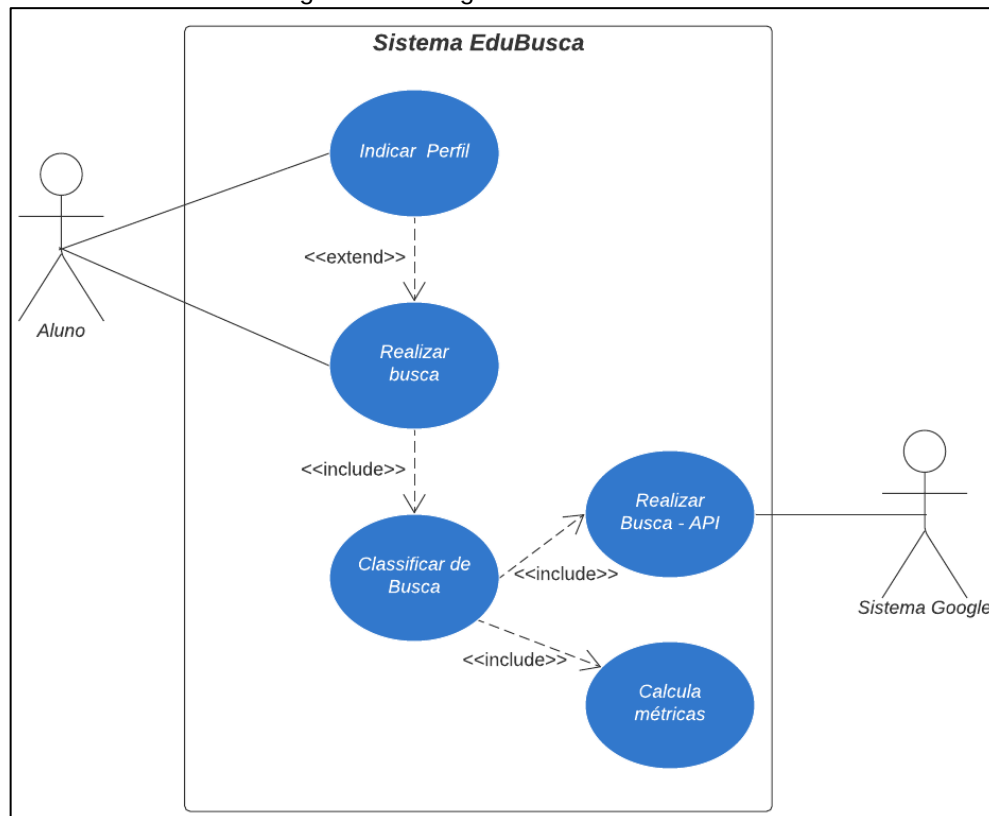
5.1.4. Motor de busca

Uma vez implementadas as metodologias de classificação dos termos de busca de modo a melhor as *strings* de busca dos alunos, o mecanismo de busca faz-se necessário para dar autonomia de pesquisa para os usuários. Na criação do ambiente de interação com usuário optou-se pela utilização da linguagem de programação Django.

O Django é uma estrutura de alto nível para programação Web, de desenvolvimento ágil, além de ser gratuito e de código aberto. Ele ainda é escrito em Python, que é a linguagem adotada na etapa de mineração de dados de dados, o que facilita no momento de interação entre as etapas do trabalho.

Para demonstrar o funcionamento do modelo implementado segue, na Figura 15, o diagrama de caso de uso.

Figura 16 – Diagrama de Caso de Uso



Fonte: Próprio Autor, 2021.

O sistema possui dois atores: Aluno e Sistema Google. O Aluno é o usuário principal do sistema, de onde partirá as solicitações do sistema. Ele pode solicitar a identificação do seu perfil inicialmente, que é o caso de uso ao qual o aluno informa qual série pertence. Deste modo torna-se possível realizar a métrica de verificação no momento do melhoramento do termo de busca, seu funcionamento é demonstrado na (Figura 18), onde é apresentada a tela de seleção do perfil.

O segundo caso de uso que pode ser solicitado pelo usuário é o de realizar a busca. Nesse é obrigatório que o caso de uso que indica o perfil tenha ocorrido, pois o valor do perfil deverá ser armazenado em uma sessão, de modo a possibilitar sua identificação durante o uso do sistema. É a partir do segundo caso de uso que o sistema classifica os termos de acordo com os identificados na base nacional comum curricular, através de inclusão com caso de uso de classificação.

Esta inclusão torna-se responsável pela inclusão do cálculo das métricas utilizadas na mineração e classificação de texto, ponto chave da pesquisa. Além da classificação, é responsável ainda por incluir o caso de uso de realizar busca via API, responsável por se comunicar com o segundo autor do sistema, o sistema do motor de busca do Google. Esse sistema, recebe o termo aprimorado que foi identificado pelas métricas durante a mineração.

O Google então recebe a *string* de busca otimizada e realiza uma busca web sobre ela. Uma vez retornado os resultados do Google, a ferramenta EduBusca lista os mesmos de acordo com a necessidade do usuário inicial. Deste modo conclui-se os ciclos possíveis da ferramenta de pesquisa, na Figura 16 é possível visualizar a tela de busca junto com a sua lista de resultados.

Figura 16 –Tela de busca do EduBusca



Fonte: Próprio Autor, 2021.

5.2 PLANO EXPERIMENTAL

Após a sistematização, o trabalho deve possuir um protocolo de aplicação, cujo objetivo é alinhar o modo de aplicar as diversas etapas da pesquisa. Para isso, definiu-se um plano experimental. Segundo Gil (2002, pg. 94), um experimento é uma pesquisa que manipula uma ou mais variáveis e com bases nessas variáveis pode-se definir os planos de experiências.

Este plano é focado na variável independente, que nesta pesquisa é: Motor de busca especializado. Para admitir a hipótese do motor de busca aperfeiçoado, proposto neste trabalho, toma-se o plano de experimento do motor de busca especializado, conforme a Tabela 4.

Tabela 4. Plano de experimento sobre motor de busca especializado

Motor de busca especializado	
A1 Melhora	A2 Indiferente
Resultados na variável dependente (avaliação dos professores pelos alunos)	

Fonte: Próprio Autor, 2021.

Para o teste da hipótese faz-se necessário a avaliação do modelo por meio de uma aplicação passível de avaliação. O plano experimental consiste em o aluno propor *strings* de buscas a serem avaliadas pelo motor de busca Google e motor de busca EduBusca. Nesta pesquisa adotou-se dois momentos de aplicação, um primeiro para teste e adaptação do modelo e o segundo momento para validar o sistema e obter os resultados do projeto junto com as conclusões.

5.2.1. Experimento de teste e adaptação

Para o experimento de teste e adaptação, que ocorreu durante a etapa do desenvolvimento, os sujeitos selecionados para a pesquisa foram 3 alunos do ensino fundamental, séries iniciais. Devido maior facilidade de selecionar alunos de uma mesma sala de aula, todos com idades e características semelhantes, com: 10 anos de idade, frequentando o quinto ano. Devido à ausência de aulas presenciais na rede de ensino, e até mesmo a distância em escolas públicas, os selecionados frequentam escola particular, em Palmas no Tocantins.

O colaborador que realizou a avaliação é um pedagogo com mestrado em história, com experiência ainda no ensino de matemática e filosofia, quando professor do ensino fundamental. Atua hoje no Instituto Federal do Tocantins, como professor do ensino médio – técnico, com a disciplina de filosofia.

Para avaliação do sistema, cada aluno escreveu uma busca, dita *string* do aluno, e a aplicação que minera (sistema) uma nova *string* de busca aperfeiçoada. Dessa, o professor colaborador realizou os testes comparativos entre os resultados das pesquisas. Com tal comparativo observou-se que a ferramenta EduBusca produz um potencial melhora nos resultados apresentados para o público.

Com esta etapa foi possível um melhor alinhamento no cálculo de relevância do motor de busca durante o processo de mineração do texto, bem como uma necessidade de atribuir um valor menor a métrica de avaliação de perfil do aluno. A redução no valor do perfil se deu porque, em alguns momentos, o aluno estar pesquisando assunto de conteúdo pertence a perfil diferente ao dele, porém próximo.

Demonstrou ainda a necessidade de um complemento para termos aos quais a BNCC não conseguia suprir o termo buscado pelo aluno (assuntos que não são abordados na base). Neste caso foi incrementado a *string* de busca do aluno o perfil ao qual ele pertencia, o que gera melhora no resultado apresentado.

É válido apontar ainda que termos aos quais os alunos apresentam erros de ortografia ou escrita não geram impactos significativos para a pesquisa. Isso se dá pelo aprimoramento da *string* que ocorre pelo processo de mineração de texto junto a BNCC. Os processos de *tokenização*, *stopwords* e *stemming*, fazem com que esses erros sejam diluídos e não impactem no prosseguimento da pesquisa na web.

Desta fase também foi possível um melhor conhecimento e preparação para a implementação da parte da interface do portal *web*. Pontos como a necessidade de verificar o perfil antes de realizar a busca e manter esta informação em uma sessão, ou ainda de conseguir mudar o perfil do aluno posteriormente surgiram da observação desta etapa da pesquisa.

Esta etapa da pesquisa foi publicada no XVIII Congresso Brasileiro de Ensino Superior à Distância (ESUD), evento pertencente a sociedade brasileira de computação e o artigo está disponível no apêndice B.

5.2.2. Experimento final

Após a experimentação de teste e adaptação o modelo ficou apto para aplicação final. Os colaboradores que participaram desta etapa foram formados por uma coordenadora pedagógica, cujo objetivo foi selecionar a turma com maior numerosidade e empenho para a participação da pesquisa e indicar os professores a colaborarem com este projeto. Houve ainda a participação de quatro professores colaboradores que compuseram a pesquisa, todos funcionários efetivos do colégio a mais de 10 anos.

Estes profissionais da educação possuem formação pedagógica com complemento em especializações na área da educação. As disciplinas a quais ministram estão contidas em matemática, português, geografia, história e educação física, mas vale lembrar que esses professores acabam atuando em algumas outras disciplinas do ensino fundamental, devido ao quantitativo de professores no colégio.

A atuação deles permitiu a observação e aplicação da pesquisa durante suas aulas. Eles auxiliaram na indicação dos temas abordados na sala de aula e classificação da turma para aplicação da pesquisa. Devido à dificuldade do andamento da pesquisa, causado pelo afastamento social gerado pela pandemia, apenas dois professores participaram do momento de avaliação dos *links*, porém, já possibilitou o cruzamento dos resultados a serem observados na conclusão desta pesquisa.

A aplicação final do projeto ocorreu em uma turma da quinta série do colégio parceiro, contando com 15 alunos no momento da interação. Os alunos tiveram 5 minutos para formular os termos de busca sobre cada um dos seguintes temas: Reciclagem, Tipos/Gêneros de poemas, Ciclo da água e Trabalho e Tecnologias. Os alunos puderam indicar a *string* de forma manuscrita pelo chat ou ainda via áudio, de forma a serem utilizadas nos motores de busca e totalizou 75 *links*, porém ao serem avaliados observou-se a duplicidade entre os alguns e após excluir estas repetições foram totalizados 33 *links* a serem avaliados pelos professores.

Uma vez relacionadas as *strings* de busca, elas foram aplicadas nos dois motores de buscas. Cada um dos motores geraram uma lista ordenada de acordo com a classificação que considera mais apta. Dentro destas listas foram elencados 5 (cinco) primeiros *links*, identificados pelo motor de busca que o gerou bem como o termo utilizado como forma de posterior identificação. O critério de seleção dos primeiros *links* foi o de que motores de busca ranqueiam de forma crescente os melhores resultados para cada pesquisa.

Após a seleção dos *links* os professores atuaram como avaliadores dos resultados selecionados. No momento da avaliação foi informado para o professor apenas o *link* a ser avaliado e a qual tema pertencia sendo oculta a informação do motor de busca que o gerou. Essa ocultação foi realizada para evitar algum tipo de influência do professor durante a avaliação dos resultados da pesquisa.

Para tal avaliação é válido entender que na geração dos resultados foi realizada a média de notas de cada uma das questões, indexadas como Q1, Q2, Q3 e Q4, seguindo a identificação do apêndice A e representa:

- Q1 – O resultado aborda o assunto desejado e necessário para o aluno;
- Q2 – O resultado possui linguagem de fácil entendimento para o aluno;
- Q3 – O resultado agrega valor ao aprendizado do aluno; e
- Q4 – O resultado pertence a faixa etária do aluno.

Foram consideradas ainda todos os formulários de avaliações dos professores, responsáveis por classificar cada *link* gerado pelos mecanismos de busca. Vale destacar que as notas consideradas no momento da avaliação seguem a escada de 0 (zero) à 3 (três), onde a menor numeração indica pior classificação no quesito avaliado.

Devido ao momento pandêmico (Covid-19) a aplicação da pesquisa ocorreu como forma de atividade durante aula remota, via *meet*. A definição da aplicação apenas com a turma do quinto ano do ensino fundamental é justificada pela dificuldade de reunir as demais turmas de forma remota durante a pandemia.

É importante destacar que ocorreram dificuldades no ensino durante a Covid-19, sobretudo para o fundamental, sendo marcadas por um longo período em que as escolas ficaram fechadas, sem qualquer tipo de interação direta com os alunos. As aulas eram realizadas de modo indireto, onde os pais buscavam o material de ensino na escola semanalmente e traziam a atividade resolvida. Devido a este modelo, esta pesquisa sofreu um lapso temporal, que só foi solucionado no final do primeiro semestre e início do segundo semestre do ano de 2021.

6 RESULTADOS E DISCUSSÕES

Neste capítulo são discutidos os resultados obtidos com a utilização das técnicas de mineração de texto aliada com a Base Nacional Comum Curricular. Essa interação possibilitou de forma satisfatória a obtenção de strings otimizadas, originadas de termos de buscas aleatório dos alunos, cujo estão em fase de desenvolvimento do pensamento cognitivo. A partir de tal conquista, torna-se possível criar resultados abordando um comparativo entre a ferramenta foco do trabalho e a principal ferramenta existente no mercado.

Para a execução do projeto contou-se com a participação dos 15 alunos, responsáveis por criar termos de buscas de acordo com os temas abordados no Item 5.2.2. Cada aluno criou um termo de busca referente a cada tema definido pelos professores e destes termos foram selecionados os 5 primeiros resultados de cada uma das ferramentas de busca para serem avaliados pelos colaboradores.

Cada ferramenta de busca gerou 165 *links* a serem avaliados, porém, por vezes os *links* se repetiam. Para evitar trabalho duplicado pelos avaliadores retirou-se as redundâncias. Desta forma, o quantitativo de *links* não duplicados pode ser observado na Tabela 5.

Tabela 5 – Quadro quantitativo de *links* por motor de busca

	Google	EduBusca
Quantidade de <i>links</i> distinto	115	57

Fonte: Próprio Autor, 2021.

A tabela anterior permite observar uma menor variação de resultados entre os alunos através do EduBusca, reduzindo os resultados em quase 65%. Esta redução já demonstra que a ferramenta consegue padronizar seus resultados tendo como base a BNCC. Por si só a redução não garante que a ferramenta tenha funcionado, mas é um indício de que houve um direcionamento dos resultados, sendo necessário a análise da qualidade dos resultados, de modo a demonstrar se de fato ocorreu melhora no conteúdo buscado.

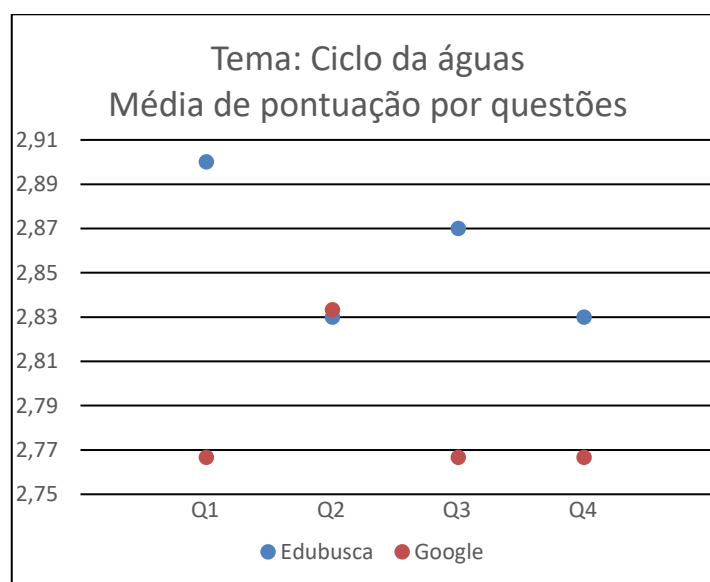
Uma vez identificados os *links*, torna-se possível a análise dos termos do mecanismo criado neste trabalho e se os objetivos foram atingidos. Para tal verificou-se os resultados das avaliações realizadas pelos professores colaboradores. Os resultados foram agrupados entre avaliações das pesquisas geradas pelo Google e

as geradas pelo EduBusca, de forma a criar um resultado quantitativo a ser analisado com visão qualitativa posteriormente.

Observando o modo de operação está detalhada no Item 5.2.2, onde cada tema de busca deve ser avaliado por quatro questões valoradas. As questões Q1, Q2, Q3 e Q4 abordam, respectivamente, se o assunto é o necessário para o aluno, se possui linguagem de fácil entendimento, se agrega valor a aprendizagem e se pertence a faixa etária dos alunos. O resultado é apresentado através da média das notas de cada questão.

A seguir são demonstrados os gráficos pontuando a média de nota, por questões, de acordo com os temas propostos no momento da busca e o motor avaliado.

Gráfico 1 – Comparativo de pontuação média do tema ciclo das águas



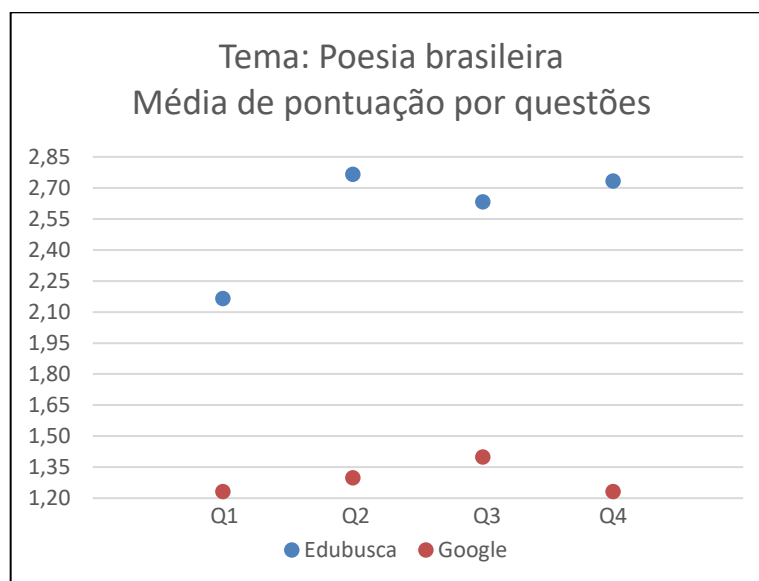
Fonte: Próprio Autor, 2021.

Quando avaliadas as questões relacionadas a: assunto desejado e necessário, que agrega valor à aprendizagem e pertencente a faixa etária (Q1, Q3 e Q4) junto ao tema ciclo da água, é possível observar o motor EduBusca atinge uma pontuação média superior 2,83 (dois vírgula oitenta e três), já o Google apresenta média inferior a 2,77 (dois vírgula setenta e sete). Assim é possível observar o ganho de desempenho no motor proposto no projeto.

Já na questão sobre o fácil entendimento para o aluno (Q2), ambos ficam com a mesma nota de avaliação. A pontuação dos dois foi boa, tendo média superior a 2,75 (dois vírgula setenta e cinco). No entanto é possível afirmar que de modo geral

o Edubusca apresenta resultados mais qualificado, possuindo média superior na temática sobre ciclo das águas, Gráfico 1.

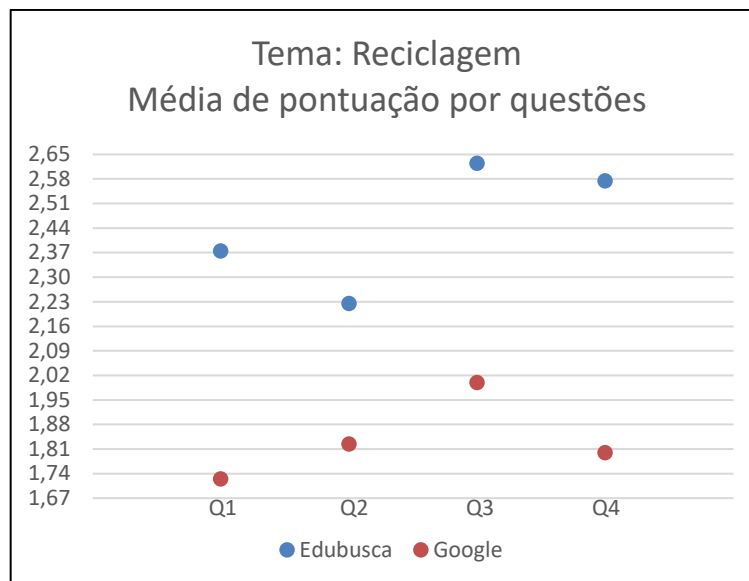
Gráfico 2 – Comparativo de pontuação média do tema poesia brasileira



Fonte: Próprio Autor, 2021.

Ao avaliar o tema da poesia brasileira (Gráfico 2), pode-se observar uma grande diferença entre as médias da avaliação entre os dois motores. O EduBusca apresenta em todas as questões uma média superior a 2,10 (dois vírgula dez). Esse resultado indica que houve melhora no tema abordado no *link* avaliado, com resultados pertencentes ao perfil, agregando valor na aprendizagem do aluno através de uma linguagem de fácil compreensão. O mesmo padrão pode ser observado nos gráficos abaixo, com os temas: Reciclagem, Tipos/Gêneros de poemas e Trabalho e Tecnologias, onde a pontuação média de todas as no Edubusca ficaram com pontuação 2,5, já para o Google apresentou média de 1,67, nota 33% menor.

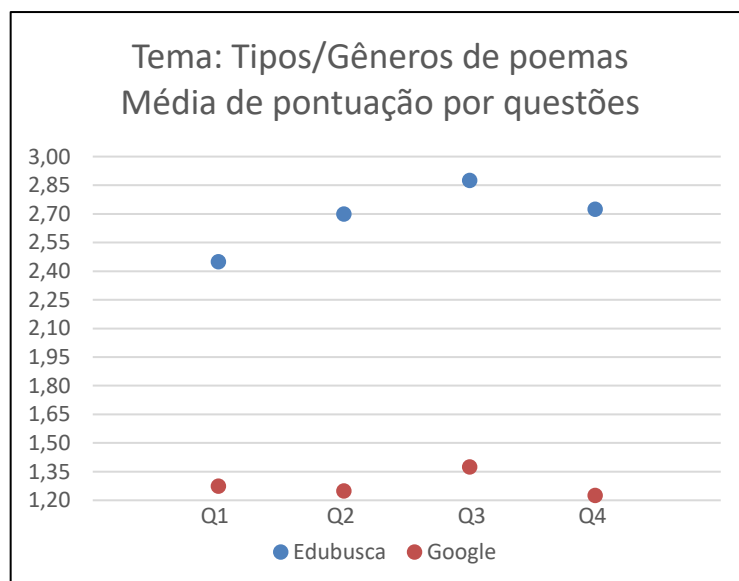
Gráfico 3 – Comparativo de pontuação média do tema reciclagem



Fonte: Próprio Autor, 2021.

Para a temática de reciclagem (Gráfico 3), a média do mecanismo EduBusca se apresentou sempre superior a 2,15 (dois vírgula quinze), o que representa uma classificação boa para o mecanismo, satisfazendo o esperado. Já o segundo mecanismo avaliado, não conseguiu em nenhuma questão atingir uma média superior a 2 (dois) uma vez que os valores ficaram entre 2 (para Q3) e 1,72 (para Q1).

Gráfico 4 – Comparativo de pontuação média do tema tipos/gêneros de poemas

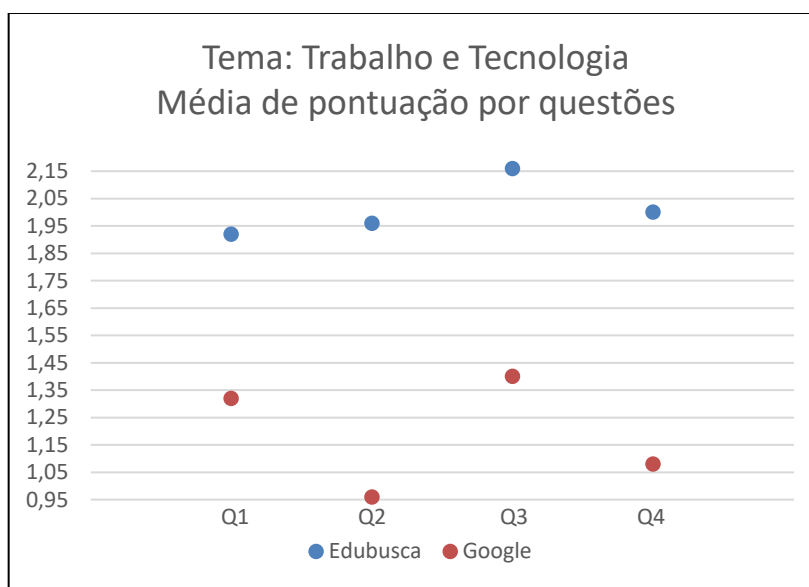


Fonte: Próprio Autor, 2021.

Para o tema relacionado a tipos e gêneros poéticos é observado um equilíbrio nas avaliações do motor Google, que apresenta uma variação na escala de

média entre 1,25 (um vírgula vinte e cinco) e 1,38 (um vírgula trinta e oito), podendo ser considerado pela avaliação como regular. Já a ferramenta desenvolvida nesta pesquisa possui uma média entre 2,45 (dois vírgula quarenta e cinco) e 2,88 (dois vírgula oitenta e oito). Assim, pode-se considerar esses resultados como ótimos, indo ao encontro do esperado para o perfil, conforme apontado no Gráfico 4.

Gráfico 5 – Comparativo de pontuação média do tema trabalho e tecnologia



Fonte: Próprio Autor, 2021.

Trabalho e tecnologia é um tema que gera grande impacto no mundo de hoje e por isso foi escolhido para ser avaliado nesta pesquisa. Assim sua avaliação (Gráfico 5) demonstra que o mecanismo EduBusca obteve as menores médias entre os demais temas. Porém, ainda assim, o resultado entre as médias foi maior que a nota 2 (dois), o que demonstra que a ferramenta teve um desempenho muito bom entre todos os temas e *strings* utilizados ao longo do projeto. Uma nota acima de 2 (dois), dentro do formulário de avaliação (Apêndice A), indica que o professor concorda que o *link* atende aquele requisito, já a nota 3 (três) indica que o avaliador concorda totalmente com o abordado pelo *link*.

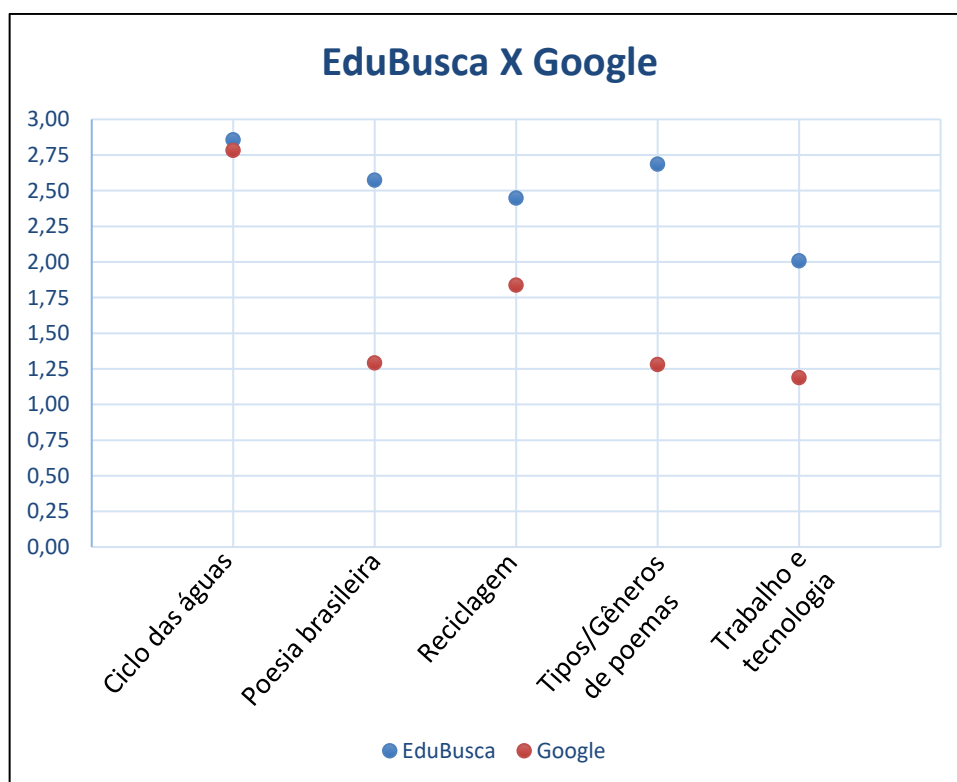
Resultados com valor 1 (um) é um resultado que possui desacordo com o desejado e 0 (zero) são resultados que estão em total desacordo com o esperado para aquele perfil de aluno. Essa classificação de parcialmente em desacordo, nota entre 1 (um) e 2 (dois), foram comuns nas avaliações do Google, informando que de modo geral a ferramenta não gera os resultados necessários para o aluno.

Como modo de melhor avaliar o desempenho como um todo deste trabalho, criou-se uma dispersão com a média global dos temas, dividindo a avaliação entre os

dois motores de busca utilizados neste trabalho (Gráfico 6). Esta avaliação indica que para o tema ciclo das águas, os motores apresentam resultados semelhantes na qualidade de retorno para as pesquisas. Quando observados os demais temas, a ferramenta aponta uma melhora significativa, como nos temas sobre poesia, cujo EduBusca retorna resultados que a nota é mais significativa para uma pesquisa. É notório ainda a média do EduBusca ser sempre maior que 2 pontos, o que demonstra uma superioridade da ferramenta.

Quando avaliado o desempenho dos motores com os demais temas é notório a dispersão entre as médias permite pontar a diferença entre a qualificação dos resultados de ambos os motores. Pode-se afirmar que o mecanismo criado neste trabalho demonstra melhor qualificação para a busca dos sujeitos analisados. Essa melhora ocorre devido ao direcionamento do EduBusca de acordo com o perfil do aluno e do documento base (BNCC), que juntos constroem o que deve ser apresentado para usuário.

Gráfico 6 – Dispersão da média global entre motores de busca por tema.



Fonte: Próprio Autor, 2021.

Expostas as validações e considerações sobre os resultados obtidos torna-se possível afirmar que esse projeto atingiu seu objetivo. A utilização das técnicas de mineração de texto aliado a BNCC fez possível criar uma ferramenta de busca, não supervisionada e que ainda assim melhora os resultados de buscas na Web para

alunos do ensino fundamental. Quando avaliado o mecanismo com a dinâmica junto à comunidade escolar, o EduBusca demonstra superioridade a principal ferramenta utilizada no mercado atual o que demonstra que este trabalho obteve êxito.

7 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho demonstra que uma ferramenta de busca que integra as técnicas de mineração de textos, aliado a Base Nacional Comum Curricular, é uma alternativa viável para criação de mecanismo de busca focado para aprendizagem do ensino fundamental. Tal técnica possibilita uma melhor adequação da *string* de busca criada pelo aprendiz através da identificação do seu perfil e alinhamento com o que direciona a BNCC.

O melhoramento dos resultados durante o comparativo entre a ferramenta criada neste trabalho e o Google é nítido e demonstra a maior adequação dos resultados quando o EduBusca é utilizado. Em todos os casos em que foram comparadas as ferramentas o mecanismo proposto obteve uma avaliação positiva, ao encontro do esperado pelos profissionais que o avaliaram, obtendo em todas as avaliações médias superiores à sua concorrente.

Durante o desenvolvimento da ferramenta, a observação as técnicas e modelos de mineração de dados foram de grande importância para a obtenção de um resultado positivo. As técnicas de classificação do resultado, levando em considerações pontos chaves como: quantidades que a palavra repete, proximidade entre palavras e verificação do perfil do aluno atingiram um equilíbrio no aperfeiçoamento do termo de busca proposto no projeto.

A mineração de texto, aplicado na BNCC, também demonstra um alinhamento com expressões a serem buscadas pelos usuários, tornando possível uma exatidão no encontro de termos desejados. Tal técnica traz ainda agilidade no momento de interação entre o usuário e o sistema *web* EduBusca. O processo de classificação de busca identifica de forma positiva os termos correlatos.

Os resultados demonstram um ganho de qualidade nos resultados das buscas pelo mecanismo proposto. Nota-se que quando o EduBusca é utilizado há um melhor direcionamento para pesquisa do usuário, podendo fazer busca realizadas com *strings* diferentes gerarem resultados iguais e mais apropriados ao perfil do usuário.

Assim, conclui-se que este projeto atingiu seu objetivo de forma satisfatória, gerando ainda uma ferramenta que pode ser testada em outros níveis da educação que tenham como base a BNCC. Aponta ainda que motores de buscas podem utilizar

técnicas de mineração de texto e se basear em documentos padrões para gerar uma melhor classificação de resultados.

Esta pesquisa traz como proposta de pesquisa futura a possibilidade de expansão do experimento para escola de ensino médio e superior, aplicando as técnicas em um documento padrão para referidos níveis educacionais. Pode-se ainda criar experimentos que utilizem a BNCC como forma de auxiliar professores a identificarem matérias que ajudem no ensino de sala de aula, principalmente ao consideram o modelo atual de aprendizado à distância.

REFERÊNCIAS BIBLIOGRÁFICAS

ARESTA, Mônica; MOREIRA, António; PEDRO, Luís. **A utilização de ferramentas Web 2.0 e a promoção de processos de aprendizagem colaborativa: implicações educativas e sociais.** *Atas do Encontro sobre Web 2.0*. Braga: CIEd. 2008.

AZAMBUJA, Marcos Jolbert Cáceres; GRIMONI, José Aquiles Baesso; CURY, Lucilene. **Educação híbrida: reflexões para a inovação.** *Opción*, v. 32, n. 7, 2016.

AZPIAZU, Ion Madrazo, et al. **Online searching and learning: YUM and other search tools for children and teachers.** *Information Retrieval Journal*, 2017, 20.5: 524-545.

BABAYIGIT, Sinan; GROPE, Sven. **Webpage ranking analysis of various search engines with special focus on country-specific search.** *Open Journal of Web Technologies (OJWT)*, v. 5, n. 1, p. 44-64, 2018.

BALDIN, Nelma; MUNHOZ, Elzira. **Snowall (Bola de neve): Uma técnica metodológica para pesquisa em educação ambiental comunitária.** *X Congresso Nacional de Educação – EDUCERE*, 2011.

BILAL, Dania; BOEHM, Meredith. **Towards new methodologies for assessing relevance of information retrieval from web search engines on children's queries.** *Qualitative and Quantitative Methods in Libraries*, 2017, 2.1: 93-100.

BRASIL. MINISTÉRIO DA EDUCAÇÃO. **Base Nacional Comum Curricular.** Brasília: MEC, 2014.

BRITO, Edeleon Marcelo Nunes de. **Mineração de Textos: detecção automática de sentimentos em comentários nas mídias sociais.** *Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento*, v. 6, n. 1, 2017.

BURGSTROM, Laura. **The Impact of Student Completion Requirements Using an LMS (Learning Management System) on Student Achievement and Differentiated**

Instruction in The Classroom. Impact of Schoology on Student Achievement. *The College at Brockport, State University of New York*, 2016, 1-39.

CABRITO, Belmiro. **Covid-19, educação (básica) e equidade em Portugal.** RTPS - Revista Trabalho, Política e Sociedade, v. 6, n. 10, p. p. 125-138, 30 jun. 2021.

CARVALHO, Ricardo César de. **Aplicação de técnicas de mineração de texto na recuperação de informação clínica em prontuário eletrônico do paciente.** 201f. Dissertação (Mestrado em Ciência da Informação) apresentada a Universidade Estadual Paulista (Unesp), Marília/SP, 2017.

CHAU, Michael, et al. **Evaluating the use of search engine development tools in IT education.** *Journal of the American Society for Information Science and Technology*, 2010, 61.2: 288-299.

Confira o ranking dos buscadores no Brasil. **Olhar Digital**, 30 de janeiro de 2015. Disponível em: <<https://olhardigital.com.br/noticia/ranking-dos-buscadores-no-brasil/46539>>. Acesso em: 06/06/2019.

CORREA, Alan; WARPECHOWSKI, Mariusa; PINTO, Andrio S. **O uso dos motores de busca na Internet: como se configuram as pesquisas de conteúdo na Web para a produção de trabalhos educacionais.** In: *Anais do Workshop de Informática na Escola*. 2014. p. 360.

COSTA, F. M. S. P. **Search Engine and Optimization.** 63f. Dissertação (Mestrado em Sistemas e Tecnologias de Informação para as Organizações) apresentada ao Instituto Politécnico de Viseu/POR. 2011.

Documentação do NLTK. NLTK, 20 de agosto de 2019. Disponível em: <<https://www.nltk.org/>>. Acesso em: 30/10/2019.

DANBY, Susan J., et al. **Going online: Young children and teachers accessing knowledge through web interactions.** *Educating Young Children: Learning and Teaching in the Early Childhood Years*, 2013, 19.3: 30-32.

FASLAH, Roni. **Model of the use of information technology in improving the ability of student.** *Jurnal Ilmiah Econosains*, 2016, 14.1: 67-81.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2002.

GOSSEN, Tatiana; KOTZYBA, Michael; NÜRNBERGER, Andreas. **Search Engine for Children: User-Centered Design.** *Datenbank-Spektrum*, 2017, 17.1: 61-67.

GOSSEN, Tatiana; NITSCHKE, Marcus; NÜRNBERGER, Andreas. **My first search user interface.** In: *European Conference on Information Retrieval*. Springer, Cham, 2014. p. 746-749.

HAHNEL, Carolin, et al. **The role of reading skills in the evaluation of online information gathered from search engine environments.** *Computers in Human Behavior*, 2018, 78: 223-234.

HUBINÁKOVÁ, Henrieta; MIKULA, Matej. **The Importance of the Internet in the Life of Students of Media Studies.** *Media Literacy and Academic Research*, 2018, 1.1: 31-42.

HUIZENGA, J. C., et al. **Teacher perceptions of the value of game-based learning in secondary education.** *Computers & Education*, 2017, 110: 105-115.

WE ARE SOCIAL, HOOTSUITE. (2021). **Digital 2021.** Retirado de, <https://wearesocial.com/digital-2021>

JOCHMANN-MANNAK, Hanna, et al. **Children searching information on the Internet: Performance on children's interfaces compared to Google.** In: *SIGIR*. 2010. p. 27-35.

KAMARUDIN, Abrar Noor Akramin; RANAIVO-MALANÇON, Bali; MUSA, Nadianatra. **Design and Implementation of PIAK: A Personalized Internet Access System for**

Kids. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 2017, 9.2-9: 49-57.

NAGY, Ádám; KÖLCSEY, Attila. **Generation Alpha: Marketing or Science.** *Acta Technologica Dubnicae*, 2017, 7.1: 107-115.

NAIR, Anjana A.; SANTHA, S. **Role of internet in education of commerce students.** *Journal Homepage: <http://www.ijmra.us>*, 2017, 7.12.

MARTINS, Diogo; et. al. **Desenvolvimento de Interfaces Digitais Infantis: Estudo Preliminar sobre Design Centrado na Criança.** *Journal Human Factors in Design*, 2017.

MOODY, Mia; BATES, Elizabeth. **PR Students' Perceptions and Readiness for Using Search Engine Optimization.** *Journalism & Mass Communication Educator*, 2013, 68.2: 166-180.

MORAIS, Edison Andrade Martins; AMBRÓSIO, Ana Paula L. **Mineração de textos.** Relatório Técnico–Instituto de Informática (UFG), 2007.

PERA, Maria Soledad; NG, Yiu-Kai. **Using online data sources to make query suggestions for children.** In: *Web Intelligence*. IOS Press, 2017. p. 303-323.

PRODANOV, Cleber Cristiano. **Metodologia do trabalho: métodos e técnicas da pesquisa e do trabalho acadêmico.** Novo Hamburgo: Feevale, 2013.

READ, Janet; MACFARLANE, Stuart. **Using the fun toolkit and other survey methods to gather opinions in Child Computer Interaction.** *Proceeding of the 2006 Conference on Interaction Design and Children*, 2006. 81-88.

REHMAN, Zobia et al. **Morpheme matching based text tokenization for a scarce resourced language.** *PloS one*, v. 8, n. 8, p. e68178, 2013.

REZENDE, Solange O.; MARCACINI, Ricardo M.; MOURA, Maria F.. **O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento**. Revista de Sistemas de Informação da FSMA. Visconde de Araújo, v. n. 7 p. 7-21, 2011.

RIEH, Soo Young, et al. **Towards searching as a learning process: A review of current perspectives and future directions**. *Journal of Information Science*, 2016, 42.1: 19-34.

ROCHA, Rafael Port. **Metadados, Web Semântica, Categorização Automática: combinando esforços humanos e computacionais para a descoberta e uso dos recursos da web**. Em Questão, Porto Alegre, v. 10, n. 1, p. 109-121, jan./jun. 2004.

RUIZ, Evandro Eduardo Seron. **Mineração de Texto em Saúde**. *Journal of Health Informatics*, 2016, 8: 1-2.

SEEMS, Patrick. **Funcionamento dos mecanismos de pesquisa**. Editora Bibliomundi, 2020.

SILVA, Sandra Regina Rocha. **O Uso de Colaboração Social na Busca para Crianças**. Tese (Doutorado em Engenharia de Sistemas e Computação) – UFRJ/COPPE. Rio de Janeiro, 2016.

SILVA, Sandra Regina Rocha; XEXÉO, Geraldo Bonorino. **Child Search Framework: A collaborative information retrieval architecture to assist children in the search process**. In: *Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on*. IEEE, 2013. p. 551-556.

SUN, Chuen-Tsai; YE, Shu-Hao; HSIEH, Hsi-Chien. **Effects of student characteristics and question design on Internet search results usage in a Taiwanese classroom**. *Computers & Education*, 2014, 77: 134-144.

VANDERSCHANTZ, Nicholas; HINZE, Annika. **How kids see search: A visual analysis of internet search engines**. In: *Proceedings of the 31st British Computer*

Society Human Computer Interaction Conference. BCS Learning & Development Ltd., 2017. p. 84.

VERMA, Disha; et. al. **A Multi-Agent Based Personalized Search Engine with Topical Crawling Capabilities**. *IUP Journal of Computer Sciences*, 2014, 8.3: 20.

VIDINLI, I. Bahattin; OZCAN, Rifat. **New query suggestion framework and algorithms: A case study for an educational search engine**. *Information Processing & Management*, 2016, 52.5: 733-752.

XAVIER, Bruno Missi; DA SILVA, Alcione Dias; GOMES, Geórgia Regina Rodrigues. **Uma arquitetura hibrida para a indexação de documentos do Diário Oficial do Município de Cachoeiro de Itapemirim**. *Transinformação*, v. 27, n. 1, p. 83-95, 2015.

APÊNDICE A - QUESTIONÁRIO FINAL OBJETIVADO

Formulário de Avaliação do Professor

Baseado no link que esta avaliando, avalie as alternativas de acordo com o seu grau de veracidade .

1. Qual link está sendo avaliado?

2. Q1 - O resultado aborda o assunto desejado e necessário para o aluno.

Marcar apenas uma oval.

- ☐ 0 – Discordo totalmente
- ☐ 1 – Discordo
- ☐ 2 – Concordo
- ☐ 3 – Concordo totalmente

3. Q2 – O resultado possui linguagem de fácil entendimento para o aluno.

Marcar apenas uma oval.

- ☐ 0 – Discordo totalmente
- ☐ 1 – Discordo
- ☐ 2 – Concordo
- ☐ 3 – Concordo totalmente

4. Q3 – O resultado agrega valor ao aprendizado do aluno.

Marcar apenas uma oval.

- ☐ 0 – Discordo totalmente
- ☐ 1 – Discordo
- ☐ 2 – Concordo
- ☐ 3 – Concordo totalmente

5. Q4 – O resultado pertence a faixa etária do aluno.

Marcar apenas uma oval.

- ☐ 0 – Discordo totalmente
- ☐ 1 – Discordo
- ☐ 2 – Concordo
- ☐ 3 – Concordo totalmente

Este conteúdo não foi criado nem aprovado pelo Google.

Google Formulários

APÊNDICE B – ARTIGO PÚBLICADO XVIII ESUD



MINERAÇÃO DE TEXTO APLICADA NA BASE NACIONAL COMUM CURRICULAR VOLTADO A OTIMIZAÇÃO DE TERMOS DE BUSCA PARA CONTEÚDO DE SERIES INICIAIS DO ENSINO FUNDAMENTAL

Laercio Pontin Junior (laerciopontin@gmail.com, Universidade Federal do Pará)

Heleno Fülber (fulber@ufpa.br, Universidade Federal do Pará)

Adonney Allan de O. Veras (allanverasce@gmail.com, Universidade Federal do Pará)

Bruno Merlin (brunomerlin@ufpa.br, Universidade Federal do Pará)

RESUMO. Ferramentas de busca *web* são poderosos auxiliares na aprendizagem infantil, porém a formação de termos de busca por crianças do ensino fundamental (séries iniciais) nem sempre retorna os resultados esperados. Alguns autores abordam temáticas semelhantes a essa, no entanto adotam um contexto de busca supervisionadas, que necessita de auxílio de pais e pedagogos para filtrar o conteúdo pertinente. Assim, este trabalho elabora um mecanismo de busca aperfeiçoado e com uma abordagem não supervisionada que, através de técnicas de mineração de texto e um documento base, consegue aprimorar pesquisas infantis na internet, gerando expressões otimizadas que apontam uma melhora dos resultados.

Palavras-chave: Motor de busca. Educação. Ensino Fundamental. BNCC.

ABSTRACT. TEXT MINING APPLIED IN THE NATIONAL COMMON CURRICULUM BASED ON OPTIMIZING SEARCH TERMS FOR INITIAL GRADE CONTENT OF ELEMENTARY EDUCATION. Web search tools are powerful assistants in children's learning, however the formation of search terms for elementary school children (initial grades) does not always return the expected results. Some authors address themes like this one, however they adopt a supervised search approach, which needs the help of parents and educators to filter the relevant content. Thus, this work elaborates an improved search engine and with an unsupervised approach that, through text mining techniques and a base document, manages to improve children's searches on the internet, generating optimized expressions that point to an improvement in the results.

Keywords: Search engine. Education. Elementary School. BNCC.

1. INTRODUÇÃO

Um motor de pesquisa ou motor de busca (*Search Engine*) é um *software* que realiza a indexação de conteúdo disponíveis na internet, mais propriamente *websites*, permitindo que os usuários dos motores pesquisem qualquer tipo de informação através de palavras-chave (COSTA, 2011).

Parte das ferramentas de buscas atuais não tem preocupação quanto ao público infantil e em seus resultados de busca, por vezes, retornam informações impróprias para a idade, além de não possuir filtro contra conteúdos adultos, propagandas e linguagens que não são habituais ao perfil do público. Sugestões de consulta podem exigir o nível de leitura avançada sobre temas complexos que crianças têm dificuldade de compreender (PERA e NG, 2017).

Existem ainda inúmeros trabalhos propondo ferramentas de busca voltadas para o público infantil, como as propostas por Silva e Xexéo (2013), Verma et. al. (2014), Vidinli e Ozcan (2016), Gossen, Kotzyba e Nürnberger (2017), entre outros. No entanto, esses propõem uma forma de pesquisa supervisionada, onde um colaborador filtra os conteúdos buscados pela criança.

Então, como obter autonomia na busca, de forma não supervisionada, de modo não apenas a filtrar conteúdos inadequados, mas também obter melhor adequação dos resultados ao público-alvo? Uma forma de automatizar a filtragem deste conteúdo é a utilização de técnicas de mineração de texto em um documento base, que rege os temas e abordagens a serem estudados pelo público.

No Brasil, o documento que rege conteúdos e matérias das mais diversas esferas da educação, é a Base Nacional Comum Curricular (BNCC), publicado em abril de 2017. Essa base é um documento de caráter normativo que define o conjunto orgânico e progressivo de aprendizagens essenciais que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica.

Para a consolidação de termos de otimização de buscas em motores *web* é importante a classificação e identificação de conteúdo pertinentes. Para isso, a literatura sugere que pais e educadores avaliem e classifiquem estes documentos. Mas, como pilar desta pesquisa é abordada uma característica não supervisionada de classificação, é utilizada a base nacional comum curricular.

A Base Nacional Comum Curricular, segundo Ministério da Educação, é:

"[...] um documento de caráter normativo que define o conjunto orgânico e progressivo de aprendizagens essenciais que todos os alunos devem desenvolver ao longo das etapas e modalidades da Educação Básica, [...], em conformidade com o que preceitua o Plano Nacional de Educação (PNE)." (MINISTÉRIO DA EDUCAÇÃO, 2017, pg. 7)

Neste documento é apontado o que os alunos necessitam aprender durante sua educação básica (direciona o que é ensinado para os alunos desta fase da educação). Nele é pontuado, de maneira clara, as temáticas por matéria, os objetos de conhecimentos das mesmas e ainda as habilidades que o aluno terá com tal conhecimento, podendo ser acessado pelo site:

<http://download.basenacionalcomum.mec.gov.br/>, disponível em diversos formatos (PDF, Planilha XLSX, entre outros).

Os objetos de conhecimentos da matéria e as habilidades, discutem sobre o que deve estar contido na matéria. Esses explicam o que deve ser ministrado sobre um assunto específico. Deste modo, eles foram utilizados para classificar a busca dos alunos de acordo o que rege a BNCC através da mineração de texto. Dele também é retirado a *string* otimizada, a qual será sugerida como melhor forma de realização de busca em motores da *web*.

2. DESENVOLVIMENTO

Segundo Gil (2002, pg. 31), após definido o problema solucionável da pesquisa, deve-se “[...] oferecer uma solução possível, mediante uma proposição, ou seja, uma expressão verbal suscetível de ser declarada verdadeira ou falsa. A essa proposição dá-se o nome de hipótese.” Definiu-se então as seguintes hipóteses:

Hipótese Nula: A *string* de busca educacional criadas por crianças em seres iniciais, quando processado por motores de buscas convencionas, não geram resultados satisfatórios ao público. (H0)

Hipótese de Pesquisa: Se a *string* de busca educacional criada por crianças em seres iniciais, quando processado por motores de buscas convencionas, não geram resultados satisfatórios ao público, então um aperfeiçoamento na *string* de busca, através de mineração de texto e da BNCC, é o ideal para buscas educacionais. (H1)

Autores, como Prodanov (2013), apontam que existem dois tipos de variáveis, que elas são “[...] as características observáveis do fenômeno a ser estudado e existem em todos os tipos de pesquisa.” Assim surge a divisão entre variáveis independentes (que influenciam e afetam as demais variáveis) e dependentes (que são influenciadas e que se busca explicar). Assim, de acordo com a hipótese, define-se:

- **Independentes:** Motor de busca convencionais e alunos;
- **Dependentes:** *String* de busca e qualidade no retorno de busca na educação infantil.

Devido à pesquisa ter sido desenvolvida em um período pandêmico (COVID-19), os sujeitos selecionados foram 3 alunos de series iniciais, com idade entre 9 e 11 anos. Eles possuem acesso a um computador com internet e possuem familiaridade com a mesma. Houve também a participação de um professor colaborador, com formação pedagógica e familiaridade na área, de forma a avaliar os retornos gerados das buscas. Sendo o responsável por avaliar os resultados, validando o modelo aqui proposto.

Para avaliação do modelo, os sujeitos elaboraram três termos de busca, de temas diferentes, que em seguida foram processados pelo mecanismo criando um padrão de busca, baseado na BNCC, através da mineração de texto. Ambos os termos de buscas foram submetidos no motor de busca Google (escolhido de forma arbitrária, por ser o mais utilizado motor na atualidade). Com retorno dos resultados da pesquisa no Google, foram selecionados os 5 primeiros links que foram avaliados pelo professor colaborador.

2.1 Implantação

Um trabalho de mineração de texto busca observar um documento padrão a ser analisado e que guie a busca assistida do modelo. Para tal, definiu-se o fluxo de trabalho de acordo com as boas práticas de mineração, que foram ilustradas na (Figura 1).

Figura 1 – Fluxograma de trabalho de mineração de texto.



Fonte: Elaborado pelo autor, com base na pesquisa realizada.

A mineração desta pesquisa utilizou a linguagem Python através da interface gráfica de usuário (IDE) Anaconda Navigator. Nesse foi utilizada a ferramenta de desenvolvimento denominada Spyder, definida por ser uma ferramenta abrangente de desenvolvimento com a exploração de dados, execução interativa, inspeção profunda e belos recursos de visualização de um pacote científico. Os códigos-fonte são públicos e gratuitos, podem ser acessados através da plataforma Github.

2.2 Pré-processamento da base nacional comum curricular

Para o funcionamento da técnica de processamento de dados, é fundamental um documento inicial de busca dos padrões a serem analisados, deste documento se abstrai a informação desejada e permite dar continuidade na abordagem estudada. Segundo CARVALHO (2017), "tais documentos são recebidos e automaticamente resumidos, internamente categorizados por um padrão de texto, e enviados automaticamente para os destinos adequados".

Nesta etapa foram retiradas do arquivo original dados sem importância para a mineração de texto, como: Cabeçalho, Rodapé, Informações Gerais do Documento. Foram removidas abas que não possuem importância por gerarem duplicidade de informações, além de retirada coluna com informações sem relevância para a mineração.

Vale destacar que além de facilitar a compreensão e implementação da mineração de texto, o pré-processamento do arquivo diminuiu o tamanho em *Kilobytes* (KB) do documento .xlsx. O documento original possui 3.188 KB, já o processado possui 645 KB. Com isso, diminui-se o tempo e o uso de recursos computacionais necessários.

2.3 Seleção de dados

O documento utilizado é uma planilha com diretrizes da BNCC, como forma de análise do arquivo é utilizada a técnica de varredura de linhas, colunas e abas. Para sua leitura utiliza-se a biblioteca WIN32COM.CLIENT que possibilita a abertura de arquivos no formato .xlsx, leitura de informações em suas diversas abas e manipulação de dados quando necessário.

Na sequência, como forma de recuperar os dados a serem processados nas rotinas de mineração de texto, utiliza-se a biblioteca NumPy, que possibilita a inserção das informações da planilha em *Arrays*, utilizados no processo de mineração de textos.

Feita a leitura da planilha, os dados são inseridos em estrutura de *Arrays* a serem processados em posterior pelas técnicas de Descoberta de Conhecimento em Dados não Estruturados. Como este trabalho visa a identificação e qualificação de expressões linguísticas, para isso ocorre uma análise semântica realizada por intermédio das técnicas de mineração.

2.4 Preparação dos dados

Recuperados os dados a serem processados, faz-se necessário o seu preparo, que consiste na aplicação de técnicas que extraem e integram os dados, realizam sua limpeza e reduzem seu volume inicial. Esse desenvolvimento é útil para a não ocorrência de informações redundantes e/ou desnecessárias. Para tal, adotou-se as técnicas conhecidas como; Tokenização, retirada de *Stopwords* e *Stemming*. Todas estas técnicas são amplamente discutidas por autores como Brito (2017) e Moraes e Ambrósio (2007).

Abaixo é apresentado o quantitativo de palavras (Tabela 1) bem como a porcentagem da numerosidade inicial e pós-processamento do arquivo de mineração deste trabalho, após aplicação das técnicas anteriores. Tamanha a diminuição faz com que se obtenha uma melhor qualidade durante a execução do trabalho.

Tabela 1 – Quadro de redução da numerosidade de palavras

	Quantidade	Porcentagem
Palavras Iniciais	276045	100%
Pós-Tokenização	39786	14,40%
Pós-Stopwords	25297	9,20%

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

Após o pré-processamento do documento, foi modelado a base de dados e realizada a carga inicial a partir da BNCC. Assim torna-se possível avaliar expressões textuais de forma estruturada, o que permite a execução da mineração de texto. Criou-se a base de dados de acordo com a (Figura 2), e adotou-se o banco de dados MySQL.

Figura 2 – Diagrama Entidade-Relacionamento.



Fonte: Elaborado pelo autor, com base na pesquisa realizada.

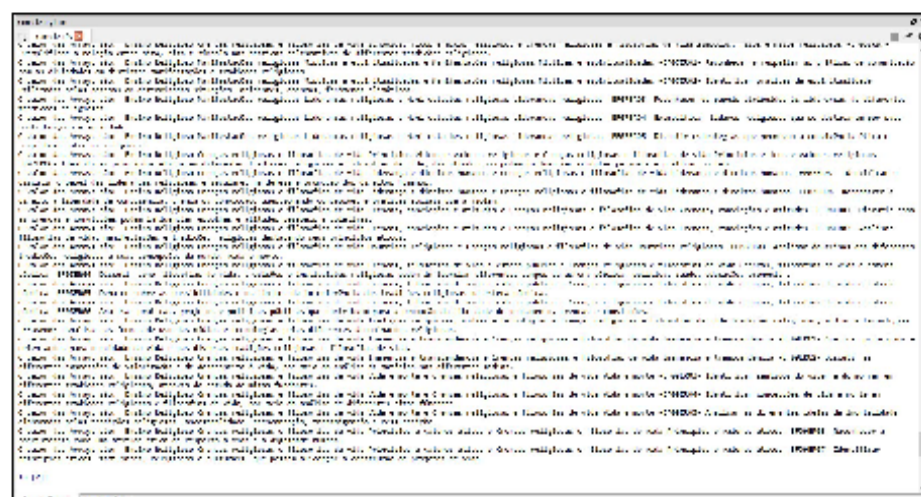
Neste trabalho, como forma de mineração de linguagem natural foi adotado o modelo de busca direta e dele gerou-se o cálculo de relevância de termos. Deste cálculo ocorre um balanceamento e os termos com maior pontuação são os considerados com retorno mais relevantes para o usuário.

Dentro do método de busca direta, os pesos foram calculados através de três métricas: frequência de vezes que a palavra é encontrada no texto, localização do termo no texto e a distância entre os termos buscados dentro do texto. Como forma de balanceamento e normalização no cálculo de relevância é realizado o balanceamento dos valores, a depender da métrica utilizada.

Todas as métricas possuem mesmo peso durante a avaliação, ou seja, 1/3 (um terço) do peso está vinculado a frequência da palavra no texto, outro 1/3 está vinculado a métrica de localização e o último 1/3 está vinculado à distância. Estes pesos foram definidos arbitrariamente e podem sofrer modificação em caso de calibragem de métricas à medida que haja necessidade. Somados os valores das métricas elas são vinculadas ao termo otimizado de forma a identificar qual o mais indicado. A *string* que possuir o maior valor de score, possui maior relevância e aproximação com o objeto de pesquisa do usuário.

Como teste de funcionalidade, demonstra-se o console do Spyder com interação que por meio de uma busca, sugere uma *string* otimizada por intermédio da mineração de textos, (Figura 3).

Figura 3 – Demonstração do uso da mineração na BNCC.



Fonte: Elaborado pelo autor, com base na pesquisa realizada.

2.5 Preparação dos dados

Os sujeitos selecionados para a pesquisa foram 3 alunos do ensino fundamental, séries iniciais. Devido maior facilidade de selecionar alunos de uma mesma sala de aula, todos possuem idades e características semelhantes, com: 10 anos de idade, frequentando o quinto ano. Devido à ausência de aulas presenciais na rede de ensino, e até mesmo a distância em escolas públicas, os selecionados frequentam escola particular, em Palmas no Tocantins. Neste trabalho utiliza-se o codinome A1, A2 e A3, para melhor identificação ao longo deste trabalho.

O colaborador que realizou a avaliação é um pedagogo com magistério em história, com experiência ainda no ensino de matemática e filosofia, quando professor do ensino fundamental. Atua hoje no Instituto Federal do Tocantins, como professor do ensino médio – técnico, com a disciplina de filosofia.

Para avaliação do sistema, cada aluno escreveu uma busca, dita *string* do aluno, e a aplicação que minera (sistema) uma nova *string* de busca aperfeiçoada. Dessa, o professor colaborador realizou os testes comparativos entre os resultados das pesquisas. Na (Tabela 2), estão contidas as *strings* criadas pelos alunos e na sequência as *strings* aperfeiçoadas pelo mecanismo.

Tabela 2 – Quadro de redução da numerosidade de palavras

Aluno: A1	Tema		
	Trabalho e Tecnologia	Porcentagem e fração	Reciclagem
<i>String</i> do Aluno	O trabalho e a tecnologia	Cálculo de porcentagem e fração	Tipos e reciclagem
<i>String</i> Proposta	Geografia mundo do trabalho, trabalho e inovação tecnológica	Matemática cálculo de porcentagens e representação fracionária	Tipos de reciclagem no ensino fundamental sem anúncio

Aluno: A2	Tema		
	Ciclo das águas	Poluição	População nômade
<i>String</i> do Aluno	Etapas do ciclo das águas	Tipos de poluição	Quem eram os nômades
<i>String</i> Proposta	Etapas do ciclo das águas - ensino fundamental sem anúncio	Geografia natureza, ambiente e qualidade de vida diferentes tipos de poluição	História povos e culturas: meu lugar no mundo e meu grupo social o que forma um povo: do nomadismo aos primeiros povos sedentarizados

Aluno: A3	Tema		
	Movimentos da Terra	Unidade de Medidas	Palavras derivadas e compostas
<i>String</i> do Aluno	Quais os movimentos da terra	Quais as unidades de medidas	O que são palavras derivadas e compostas
<i>String</i> Proposta	Ciências terra e universo forma, estrutura e movimentos da terra	Matemática grandezas e medidas	O que são palavras derivadas e compostas no ensino fundamental sem anúncio

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

Nessa tabela, pode-se observar algumas características das formulações de busca do perfil estudado. Suas buscas são derivadas do tema a ser pesquisado, acrescentando algumas palavras chaves, sem uma maior contextualização. Com isso, gera-se no retorno de busca alguns artigos e conteúdo que não atendem seu perfil, com linguagem incompatível e, em alguns casos, anúncios.

Observa-se ainda que em algumas questões não foram retornadas *strings* propostas derivadas da BNCC. Com isso, o sistema repete a *string* do aluno e acrescenta informações direcionando a busca para matérias de ensino fundamental e sem

anúncio. Com estas informações o retorno das pesquisas apresenta melhora nos resultados, apesar de uma forma simples no aperfeiçoamento da *string* do aluno.

Assim, é possível identificar ainda que a BNCC não consegue indicar todas as pesquisas que os alunos podem buscar e desta forma deve-se criar soluções que auxiliem a busca, sem a utilização de informações da BNCC. A existência de outros documentos a serem minerados, como forma de complementar os dados do documento utilizado nesta pesquisa, podem gerar resultados com maior aprofundamento.

Como forma de avaliação dos resultados, adotou-se 4 perguntas, que foram analisadas a cada resultado das *strings* avaliadas, são elas:

- Q1 – É o assunto desejado pelo aluno?
- Q2 – Possui linguagem de fácil entendimento?
- Q3 – Agrega valor ao aprendizado?
- Q4 – Pertence a faixa etária do aluno?

As respostas para as questões foram também definidas, para que se possa avaliar de forma quantitativa o resultado do trabalho, sendo adotado as seguintes opções de avaliação, por notas:

0 – Para resultados considerados sem nenhum vínculo com o perfil do aluno, que apresentam propaganda ou não se encaixa de forma alguma com o que o aluno necessita;

1 – Para resultados pouco relevante, que possui pouco grau de vínculo com entendimento do aluno;

2 – Para resultados com relevância, onde é abordado o assunto correto para a busca, de modo assertiva, mas que ainda possui algum grau de divergência com aluno;

3 – Para resultados com grande relevância, abordando perfeitamente o que é de interesse do aluno.

Para avaliar, foram considerados apenas os cinco primeiros resultados apresentados pelo Google, como descrito no Item 3. Seguem as tabelas 3 e 4, que apresenta a avaliação de apenas uma única *string* de busca, como forma demonstrativa do funcionamento da avaliação realizada pelo colaborador. A *string* de busca demonstrada pertence ao primeiro aluno (A1), onde a Tabela 3 pontua os valores para *string* de busca do aluno e a 4 a pontuação após utilização do mecanismo de melhoramento.

Tabela 3 – Avaliação da *String* de busca do aluno A1

<i>String</i> do A1: Trabalho e tecnologia	Q1	Q2	Q3	Q4
Resultado 1	0	0	0	0
Resultado 2	1	0	0	0
Resultado 3	3	2	3	2
Resultado 4	3	2	3	2
Resultado 5	2	1	2	1

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

Tabela 4 – Avaliação da *String* de busca do mecanismo, relacionado a Tabela 3

<i>String</i> do mecanismo baseado na tabela 3: Geografia mundo do trabalho e inovação tecnológica	Q1	Q2	Q3	Q4
Resultado 1	1	1	1	1
Resultado 2	3	3	3	3
Resultado 3	1	1	2	1
Resultado 4	3	2	3	2
Resultado 5	1	1	2	1

Fonte: Elaborado pelo autor, com base na pesquisa realizada.

No Gráfico 1, é possível ver a média de avaliação geral entre todas as buscas, nela é demonstrada a somatória de todas as avaliações de uma questão e retirada a sua média (soma-se todos os valores da questão e divide pela quantidade de vezes que ocorreram). Nessa, é feita apenas a divisão entre as questões que formam formuladas pelos alunos e as que foram propostas pelo mecanismo.

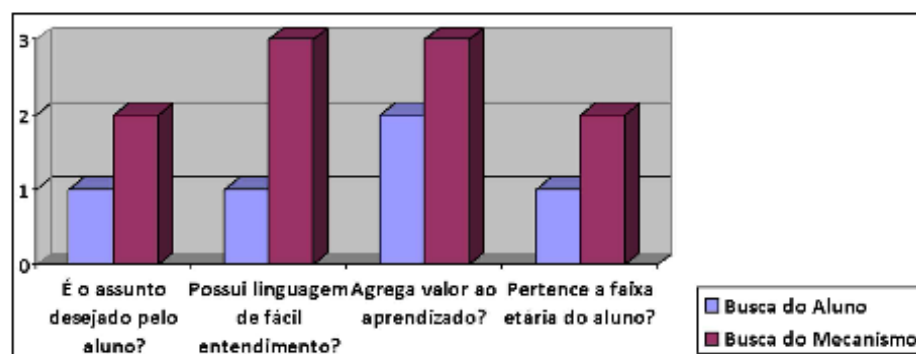


Gráfico 1. Média de qualidade na avaliação global da busca do aluno e mecanismo.
Fonte: Elaborado pelo autor, com base na pesquisa realizada.

Através destas duas medidas é possível afirmar que houve ganhos nos resultados, segundo avaliação do professor colaborador dentro do ambiente da pesquisa. É notório que todos os pontos avaliados obtiveram ganho de qualidade no retorno da busca, o que demonstra um potencial sucesso na utilização do mecanismo proposto em maior escala.

3. CONCLUSÃO

Este trabalho apresentou que a mineração de textos, aliado a BNCC, possibilita a adequação de *strings* de pesquisa em motores de buscas. Foram realizados testes que demonstraram perceptível melhora nos resultados após utilização da técnica aplicada. Possibilitando visualizar uma melhora potencial nas buscas de conteúdo para alunos das séries iniciais do ensino fundamental.

O melhoramento dos resultados, como demonstrado, tornam as buscas mais adequadas ao público. O mecanismo se comporta como aperfeiçoamento de *string* de busca, conseguindo gerar ganhos durante o uso pelo público. Observa-se ainda que todos os pontos avaliados, quando utilizado a busca do mecanismo, estão com a média de consideradas "Relevantes" ou "Muito Relevante", o que demonstra uma alta qualificação dos resultados, podendo salientar que em nenhum desses resultados geraram conteúdos indesejados, como anúncios.

Entretanto, com a pandemia COVID-19, tais testes não puderam ser realizados em grande escala ou com alunos de escolas públicas. É sabido que escolas públicas, em sua maioria, possui alunos com menor renda e consequentemente, com menos contato com a tecnologia. Então, este público também deve ser avaliado ainda na proposta deste trabalho.

Conclui-se assim, de maneira preliminar, ser possível realizar o melhoramento de termos de busca através do uso de mineração de texto no documento padrão da Base Nacional Curricular Comum, gerando expressões otimizadas que retornam termos reelaborados com potencial de melhora dos resultados. Assim, termos elaborados com palavras técnicas apresentaram melhores resultados.

A adoção de um auxiliar para buscas que não podem serem mineradas pela BNCC deve ser avaliada e implementada, pois algumas buscas não podem serem melhoradas pelos dados do documento. Contudo, este trabalho possui um resultado positivo e aponta que a adoção de mineração de texto no documento padrão e oficial do ministério da educação pode satisfazer e auxiliar alunos das diversas séries do ensino fundamental a terem um resultado mais pertinente em suas buscas relacionadas a aprendizagem escolar por intermédio da internet.

Esta pesquisa encontra-se em estágio parcial, devido ao período pandêmico, o que permite propor como trabalho futuro a ampliação da discussão dos achados de forma que confronte os resultados com as metodologias ativas de ensino. Tal feito torna possível a validação dos resultados desta pesquisa de forma macro. Pode-se ainda implementar o mecanismo proposto atuando em outros documentos padrões da educação e verificar a efetividade perante os expostos nesta pesquisa.

REFERÊNCIAS

- BRASIL. MINISTÉRIO DA EDUCAÇÃO. (2014) "Base Nacional Comum Curricular", Brasília.
- BRITO, E. M. N. (2017) "Mineração de Textos: detecção automática de sentimentos em comentários nas mídias sociais". Projetos e Dissertações em Sistemas de Informação e Gestão do Conhecimento, v. 6, n. 1.
- CARVALHO, R. C. (2017) "Aplicação de técnicas de mineração de texto na recuperação de informação clínica em prontuário eletrônico do paciente", In: Dissertação (Mestrado em Ciência da Informação) apresentada a Universidade Estadual Paulista (Unesp).

COSTA, F. M. (2011) "Search Engine and Optimization", In: Dissertação (Mestrado em Sistemas e Tecnologias de Informação para as Organizações, Instituto Politécnico de Viseu/POR.

GIL, A. C. (2002) "Como elaborar projetos de pesquisa". 4. ed. São Paulo: Atlas.

GOSSEN, T.; KOTZYBA, M. E NÜRNBERGER, A. (2017) "Search Engine for Children: User-Centered Design". Datenbank-Spektrum, 17.1: 61-67.

MORAIS, E. A. M. E AMBRÓSIO, A. P. L. (2007) "Mineração de textos. Relatório Técnico-Instituto de Informática" (UFG).

PERA, M. e NG, Y. (2017) "Child Search Framework: A collaborative information retrieval architecture to assist children in the search process", In: Web Intelligence, IOS Press.

PRODANOV, C. C. (2013) "Metodologia do trabalho: métodos e técnicas da pesquisa e do trabalho acadêmico". Novo Hamburgo: Feevale.

SILVA, L. A. (2016) "Introdução à mineração de dados: com aplicações em R." Elsevier Editora Ltda.

SILVA, S. R. R. e XEXÉO, G. B. (2013) "Child Search Framework: A collaborative information retrieval architecture to assist children in the search process." In: Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on. IEEE, p. 551-556.

VERMA, D. et. al. (2014) "A Multi-Agent Based Personalized Search Engine with Topical Crawling Capabilities", In: IUP Journal of Computer Sciences.

VIDINLI, I. B. E OZCAN, R. (2016) "New query suggestion framework and algorithms: A case study for an educational search engine." Information Processing and Management, Volume 52, Issue 5, Pages 733-752

APÊNDICE C – ARTIGO PÚBLICADO XVIII WEIBASE

Ferramentas de Busca para Crianças: revisão sistemática da literatura

Laercio Pontin Junior, Heleno Fülber, Adonney Allan de Oliveira Veras, Bruno Merlin

Programa de Pós Graduação em Computação Aplicada – Universidade Federal do Pará (UFPA)

Tucuruí – PA – Brasil

{laerciopontin,fulber,allanverasce,bruno.merlim}@gmail.com

Abstract. *This article describes the systematic review of the literature on the definition of how web search tools are used today by students of the initial series, as well as their characteristics and faults. To do this, we search for references in existing bibliographies, through the definition of key words and search string, which, in line with filtering the results of journals with the Snowboll technique, results in the selection of articles relevant to the topic. These make it possible to identify that there is research focused on search engines for children, but note that it is still an open topic, if observed its complexity.*

Resumo. *Este artigo descreve a revisão sistemática da literatura voltada para a definição de como as ferramentas de buscas Web são utilizadas hoje por alunos de series iniciais, bem como quais são suas características e falhas. Para tal, busca-se referencias nas bibliografias existentes, através de definição de palavras chaves e string de busca, que, alinhado com filtragem de resultados de periódicos com a técnica Snowboll, resulta na seleção de artigos relevantes ao tema. Esses possibilitam identificar que existem pesquisas voltadas para motores de buscas infantis, porém, notasse que ainda é um tema em aberto, se observado a sua complexidade.*

1. INTRODUÇÃO

Grande parte das ferramentas de buscas atuais não tem preocupação com o perfil do público infantil e em como realizam buscas. Por vezes trazem informações impróprias para a sua idade, bem como não possuem filtros contra conteúdos adultos, propagandas, e linguagens que não são habituais para idade. Sugestões de consulta podem exigir o nível leitura avançadas sobre temas complexos que as crianças têm dificuldade de compreender [Pera e Ng, 2017]. Assim, ao realizar buscas em ferramentas populares, crianças estão à mercê de toda e qualquer informações contidas na *Web*.

Quando se fala em motor de busca associa-se, por muito, ao Google. Como exemplo, na Alemanha, o Google tem participação de 86,59% do mercado de motores de busca, enquanto a ferramentas Bing tem 9,20% e Yahoo possui 2,63% [Babayigit e Groppe, 2018]. Devido a esta popularidade, torna-se implícito que grande parcela de alunos, que buscam aprendizado por meio de tais ferramentas, utiliza a plataforma do Google como forma de buscar conhecimento.

Pera (2017) demonstra que as ferramentas existentes não estão trazendo um ganho considerável ao público infantil. Ao realizar uma busca rápida nos motores existentes, deparasse com motores de *interfaces* confusas, que retornam pesquisas de publicidades, ou ainda funcione com buscas limitadas. Para Verma, Minocha e Kochar (2014) até então a

sociedade não possui motor de pesquisa educacional dedicado que exhibe os resultados apenas de conteúdo educacional.

Esta revisão sistemática visa demonstrar o que existe na literatura referente ao comportamento dos motores de busca para crianças, quais as dificuldades e necessidades deste público e ainda o que há de incorreto a ser considerado em propostas futuras. Este artigo tem por objetivo responder a seguinte questão de pesquisa: “Como funciona um motor de busca que pretende atender a necessidade de aprendizagem de crianças?”.

Para complementar a questão anterior, foram derivadas as seguintes questões:

- Q1 – Quais motores de busca atendem crianças?
- Q2 – Quais características são abordadas nos motores para o público-alvo?
- Q3 – Quais as falhas nos motores de busca existentes?

2. DEFINIÇÃO DA BUSCA

Com intuito de responder à questão da revisão sistemática as seguintes palavras chaves foram definidas, para a realização de pesquisa nas bibliotecas digitais:

- *Children*;
- *Kid*;
- *Search Engine*;
- *Education*; e
- *Learning*.

A *string* de busca definida para a busca inicial, a partir das palavras chaves e que trouxe um resultado considerável satisfatório foi:

(“*Kid*” OR “*Children*”) AND (“*Search engine*”) AND (“*Education*” OR “*Learning*”)

Vale destacar que, além da definição e elaboração de *string* de busca, adotasse a técnica conhecida como *Snowball*, que é uma forma de amostra não probabilística utilizada em pesquisas sociais onde participantes iniciais indicam novos participantes (Baldin; Munhoz, 2011). Essa mesma técnica pode ser abordada em trabalho escrito, onde através de citações e referencial, é possível identificar trabalhos correlacionados e com similaridade. Assim foi gerada uma interação entre trabalhos correlacionados, o que possibilita a expansão do número de referencial considerado ao longo deste trabalho.

3. FONTES DE DADOS

Definiu-se como fonte de dados, livros, revistas, artigos, teses, dissertações, jornais e leis. A pesquisa considera a literatura publicada a partir do ano de 2013. As buscas foram realizadas na base de periódicos da Capes e *Scholar Google*, sendo selecionados trabalhos escritos em língua portuguesa e inglesa.

4. SELEÇÃO DO ESTUDO

Como a *string* de busca gerou uma quantidade expressiva de retorno que não eram pertinentes ao estudo, foram adotados critérios de inclusão e exclusão dos artigos, onde foram considerados apenas:

- Trabalhos direcionados a otimização de motores de busca;

- Trabalhos que transmite a padrões de pesquisas realizadas por alunos na educação inicial;
 - Trabalhos relacionados a mecanismo de busca otimizado para estudante;
 - Trabalhos relevantes a buscas feitas por estudantes; e
 - Quando abordado características que identificam o uso de crianças em mecanismos de busca.
- Quanto ao critério de exclusão, foi observado:
- Trabalhos que não tenham métodos claros;
 - Trabalhos cujo título difere do tema pretendido;
 - Trabalhos que não tenha motor de busca ou tecnologia na educação como tema central; e
 - Trabalhos duplicados.

4.1. Qualidade dos Estudos

Os temas que têm como foco de estudo a utilização de motores de busca correlacionados a utilização por crianças são os buscados neste trabalho. Ao realizar a busca da *string* de busca, foram retornados mais de 3.400 trabalhos, não sendo possível realizar melhoramento da mesma, pois trouxe como consequência a perda de trabalhos relevantes, como o: *PR Students' Perceptions and Readiness for Using Search Engine Optimization*.

Vale ressaltar que no retorno, houve casos de retornos que não trouxeram nenhuma das expressões da *string* de busca, sendo esses desconsiderados imediatamente. Assim, foi iniciado o refinamento dos trabalhos utilizando os critérios de inclusão e exclusão, definidos no Item 4, seguindo os filtros apontados na sequência.

O primeiro filtro realizado nos artigos foi à exclusão dos trabalhos duplicados e leitura de títulos. Tais etapas ocorreram em conjunto, pois o título caracteriza a existência de trabalhos idênticos, porém que se encontra em base de dados distinta. A leitura do título permite ainda identificar trabalhos que não atende aos critérios de inclusão e exclusão de artigos para seleção de estudo.

Aplicando ambas as técnicas, permaneceram selecionados, aproximadamente, 1.000 artigos. Realizou-se então a fase de leitura de título como eliminação, reduziu de forma considerável os textos, resultando em 30 artigos relevantes. Esta etapa foi importante, para identificar e definir o critério de exclusão de trabalhos que não possuam como temática: motores de busca ou tecnologia na educação, como tema central do artigo.

Após análise do título, selecionou-se 30 artigos, que foram lidos na íntegra, como forma de filtragem final. Nesta etapa, foram selecionados 12 artigos aos quais atenderam a necessidade desta pesquisa. Junto com a leitura da íntegra dos textos, foi possível aplicar a técnica *Snowball*, que permitiu a leitura de trabalhos citados, pelos 12 artigos selecionados e relacionados ao tema. Com tal técnica, incorporou-se aos resultados 3 artigos, totalizando 15 artigos, abordando a temática desejada.

5. Sumarização e Quadro Geral de Pesquisas

Como forma de elencar e demonstrar os trabalhos relevantes identificados, segue quadro discricionários das obras consideradas.

Tabela 1. Quadro de identificação de obras

Título do artigo	Autor(es)	Ano	Veículo
<i>Child Search Framework: a collaborative information retrieval architecture to assist children in the search process</i>	Silva, S. R. R. e Xexéo, G. B.	2013	<i>IEEE 17th International Conference on Computer Supported Cooperative Work in Design</i>
<i>A Multi-Agent Based Personalized Search Engine with Topical Crawling Capabilities</i>	Verma, D.; Minocha, K.;e Kochar, B.	2014	<i>IUP Journal of Computer Sciences</i>
<i>PR Students' Perceptions and Readiness for Using Search Engine Optimization</i>	Moody, M.;e Bates, E.	2013	<i>Journalism & Mass Communication Educator</i>
<i>Model of the use of information technology in improving the ability of student</i>	Faslah, R.	2016	<i>Jurnal Ilmiah Econosains</i>
<i>Search Engine for Children: User-Centered Design</i>	Gossen, T.; Kotzyba, M.;e Nürnberger, A.	2017	<i>Datenbank Spektrum</i>
<i>The role of reading skills in the evaluation of online information gathered from search engine environments</i>	Hahnel, C; Goldhammer, F.; Krohne, U.; Naumann, J.	2018	<i>Computers in Human Behavior</i>
<i>Effects of student characteristics and question design on Internet search results usage in a Taiwanese classroom</i>	Sun, C.; Ye, S.;e Hsieh, H.	2017	<i>Computers & Education</i>
<i>BASE (Bielefeld Academic Search Engine)</i>	Bäcker, A.; Pietsch, C.; Summann, F.;e Wolf, S.	2017	<i>Datenbank-Spektrum</i>
<i>New query suggestion framework and algorithms: A case study for an educational search engine</i>	Vidinli, I. B.;e Ozcan, R.	2016	<i>Information Processing and Management</i>
<i>Online searching and learning: YUM and other search tools for children and teachers</i>	Azpiazu, I. M.; Dragovic, N.; Pera, M. S.;e Fails, J. A.	2017	<i>Information Retrieval Journal</i>
<i>Going online: Young children and teachers accessing knowledge through web interactions</i>	Danby, S.	2013	<i>Educating Young Children: Learning and Teaching in the Early Childhood Years</i>

<i>Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries</i>	Bilal, D.; e Boehm, M.	2013	<i>Qualitative and Quantitative Methods in Libraries</i>
<i>Using online data sources to make query suggestions for children</i>	Pera, M. S.; e Ng, Y.	2017	<i>Web Intelligence</i>
<i>How Kids See Search: A Visual Analysis of Internet Search Engines</i>	Vanderschantz, N.; e Hinze, A.	2017	<i>BCS Learning & Development</i>
<i>Design and Implementation of PIAK: A Personalized Internet Access System for Kids</i>	Noor, A.; Kamarudim, A., Ranaivo-Malançon, B.; e Musa, N.	2017	<i>Journal of Telecommunication, Electronic and Computer Engineering (JTEC)</i>

Fonte: Próprio autor, 2019.

6. RESULTADOS E DISCUSSÕES

Após concluir as etapas de planejamento, execução e sumarização dos trabalhos, buscou-se a leitura dos trabalhos selecionados de forma exploratória, a fim de discutir sobre as questões definidas na fase de planejamento. A principal dificuldade encontrada na revisão foi à quantidade de trabalhos com algum tipo de envolvimento com motores de busca, isso gerou uma grande massa de dados a ser analisado.

Analisado os trabalhos e definido os de maiores relevâncias, tornou-se possível discutir cada questão elaborada anteriormente.

6.1. Q1 - Quais motores de busca atendem crianças?

As análises dos trabalhos, desta questão, apontam que existem projetos direcionados a motores de buscas com propósito infantil. No entanto, Silva e Xexéo (2013, pg.1) apontam que os atuais sistemas de colaboração remota foram criados para um público geral, de modo a atender a maior quantidade de pessoa. Com isso, ele acaba não sendo focado na necessidade específica das crianças. Deixa ainda de levar em conta as características do público infantil e suas peculiaridades.

As propostas de motores existem como é o caso do *Personalized Internet Access System for Kids* – PIAK, que foi desenvolvido por Abrar Noor Akramin Kamarudim, Bali Ranaivo-Malançon e Nadianatra Musa (2017). Essa proposta busca filtrar em duas etapas, a primeira elimina páginas indesejadas, por meio de lista negra e algoritmo. A segunda filtragem ocorre de forma colaborativa e baseada em conteúdo para recomendação, através de uma página principal segura (DMOZ).

O PIAK identifica o usuário através de um registro no sistema, e por meio desse mecanismo, verifica o conteúdo pertencente ao seu perfil (através da idade). Este sistema leva em conta ainda, o grau de escolaridade dos pais, para conseguir avaliar se o usuário poderá contar com auxílio dos pais durante a pesquisa. Ele considera ainda interface de interação entre o usuário e o sistema, adotando o tema da página do *YouTube* e a tecnologia *bootstrap*.

O modelo adotado por esta ferramenta aborda aspectos onde o usuário precisa estar com uma linguagem atendida pela DMOZ, pois ela é página principal, que será varrida pelo robô de busca. Além disso, o usuário precisa cadastrar o usuário, para que ele consiga usar o sistema, tendo a necessidade de identificar a qualificação dos pais para auxiliar a busca (uma busca assistida).

Existe ainda estudo que desenvolve o motor através do modelo de pesquisa social da arquitetura integra pais e filhos e visa melhorar o processo de pesquisa para as crianças. É o caso do *Child Search Framework* - CSF, elaborado por Sandra Regina Rocha Silva e Geraldo Bonorino Xexéo (2013). Ele aborda a utilização de busca assistida, onde o pai monitor e decide quais os melhores resultados a serem apresentado para busca de seus filhos.

Esta proposta colaborativa é realizada por meio de remoto, para assim criar um aumento da cobertura do espaço de informação e diminuindo na redundância dos trabalhos. Aliado, é proposto ainda um agente inteligente, de modo a conseguir identificar padrões de escrita e definir o que está sendo buscado.

Disha Verma, Kanika Minocha e Barjesh Kochar (2014), propõem um motor de busca multi-agente, que personalize os resultados com base no histórico de navegação do usuário e perfil explícito criado por ele. Em seu trabalho, apresentam a dificuldade dos motores de buscas atuarem em um ambiente não controlado e direcionado a conteúdo educacional e a dificuldade de identificar o resultado desejado para este tipo de busca.

É proposto que as buscas iniciem em sites que com domínios específicos para educação, e ainda que seja permitido ao usuário escolher qual tipo de local a busca ocorra (blog, site de notícia, etc..). Foca ainda no retorno de informações baseados ao histórico de atividades do usuário na Internet, utilizando para está dado no ranque dos retornos. Por fim é destacada a necessidade de uma interface amiga (*view friendly*), preocupada com o dispositivo utilizado pelo usuário (celular/*desktop*).

Já Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera e Jerry Alan Fails (2017), avaliam motores de busca existentes, voltados para a educação, como o YUM. Na identificação da busca de uma criança, abordasse a necessidade adaptativa que o motor de busca deve ter para identificar possíveis erros de escrita ou de termos impróprios. Assim, o motor retorna ao usuário possíveis formas da string de pesquisa que interesse na busca.

Outra característica imprimida no YUM, que ganha destaque é a opção de consultas sugeridas, onde uma ferramenta on-line gera palavras-chave, dando alternativas de consultas para o usuário (espécie de auto-completar). A legibilidade é empregada ainda no trabalho, fazendo os retornos adequados de acordo com a capacidade de leitura das crianças.

Deste modo são notórios os rumos em que as pesquisas voltadas ao tema se encontram, bem como sua relevância. Inferisse que a grande maioria dos mecanismos necessitam de alguma forma, identificar o perfil do pesquisador. A preocupação com a parte da interface é ainda uma constante, pois sendo amigável, torna a interação com usuário algo mais agradável.

Características que adéquem à escrita e a leitura das crianças são destaque. Porém, nenhum dos mecanismos busca um melhoramento do filtro, através de um documento base, que auxilie o mecanismo de busca a identificar o estudante deseja. Assim, esta proposta tornasse relevante, por não ter sido explorada e implementada até o momento.

6.2. Q2 - Quais características são abordadas nos motores para o público-alvo?

O estudo leva a uma série de características relevantes ao se desenvolver uma ferramenta de busca para o público. Porém, o que observasse é a necessidade de trabalhar de maneira a levar

em conta a característica do público, bem como sua imaturidade. Chuen-Tsai Sun, Shu-Hao Ye e Hsi-Chien Hsieh (2017), em seu trabalho demonstra o modo que crianças formulam suas pesquisas, propondo até ensiná-los a formular de modo apropriado suas buscas.

Tatiana Gossen, Michael Kotzyba e Andreas Nürnberger (2017), que apontam ainda a necessidade de trabalhar, com crianças, em ambientes onde o usuário seja o centro, com um design fácil e divertido, proporcionando uma experiência agradável. No trabalho é proposto um desenvolvimento que leve em conta a capacidade cognitiva, conhecimento e forneça apoio emocional para as crianças. A pontuar o requisito de suporte emocional, é apontado que crianças podem ficar frustradas rapidamente. Assim, é sugerido gerar sensações de sucesso sempre que possível e uma alternativa adequada para apontar possíveis falhas do usuário.

Leva-se em consideração o suporte de línguas, pois crianças, em muitos casos, ainda possuem dificuldade na escrita e leitura, o que pode gerar erros ortográficos. Além disso, pode haver dificuldade com o uso de teclado, outro fator de frustração da criança. Assim, é indicado que o mecanismo permita apontar correções ortográficas e sugestão de consultas ao usuário. Pode, ainda, ter como complemento, um menu de navegação, que auxilia o público que ainda não possui facilidade em formular perguntas.

Aborda-se o apoio de uma interface gráfica apoiada às habilidades das crianças, de modo que o uso seja fácil e intuitivo. O suporte ao julgamento de relevância dos sites recuperados é pontuado nesta literatura, devido a dificuldades de análise e identificação do público. Esse ponto vai de encontro com a necessidade da diversidade entre crianças, que podem ser influenciados pelas habilidades cognitivas, motoras, diferença de idades e indivíduos.

Ao observar o modo de se ranquear os motores de busca, Dania Bilal e Meredith Boehm (2013), analisam o modo que se trabalham os motores de busca, e apontam ainda que existam melhorias a serem feitas por eles. Nessa pesquisa, os autores apontam a necessidade um meio de classificação das consultas on-line, chegando a sugerir que sejam mediados por especialistas em informação, professores e/ou pais, de modo a julgar a relevância.

Nicholas Vanderschantz e Annika Hinze (2017), foca em identificar erros de buscas e quais abordagens são utilizadas nos motores, propondo melhorias em seu desenvolvimento. Entre elas está a utilização de técnicas de auxílio a construção de consultas. Assim, produz mais resultados, qualificados, e com efeito positivo por reduzir número de publicidade e informações indesejadas.

É apontada ainda a necessidade apoio as crianças por meio de diretrizes de design. Assim como em outros trabalhos, neste ponto se apresenta a necessidade de auxílio à reformulação a consulta. Indica ainda a possibilidade de inserir barras laterais que contêm informações relacionadas à consulta, que pode auxiliar os jovens pesquisadores.

6.3. Q3 – Quais as falhas nos motores de busca existentes?

Ao analisar as duas questões anteriores, junto com uma breve pesquisa nos motores especialistas para crianças, é notório falhas. Autores como Carolin Hahnel, Frank Goldhammer, Ulf Krohne e Johannes Naumann (2018), identificam que nos motores de buscas é exigida uma habilidade de leitura, para a seleção correta de resultados, pois não trazem por vezes os resultados desejados de forma direta.

Isso se dá pela necessidade de reconhecimento de palavras, compreensão semântica e compreensão da leitura, que gera a capacidade de verificar e descartar informações irrelevantes, bem como saber se informações adicionais de uma pesquisa devem ser levadas em consideração na consulta. Destacando que na educação, este déficit é ainda maior, sendo mais prejudicial.

Chuen-Tsai Sun, Shu-Hao Ye e Hsi-Chien Hsieh (2017), parte do princípio que a formulação da pergunta é uma característica fundamental. Como comprovação treina o público infantil em como realizar buscas de forma mais precisa. Destaca ainda que o aluno que possui algum tipo de motivação, como um colega para realizar consultas em dupla, gera uma melhora nos resultados obtidos no mecanismo de busca.

Ion Madrazo Azpiazu, Nevena Dragovic, Maria Soledad Pera e Jerry Alan Fails (2017), analisa ferramentas focadas na busca de crianças e indica a necessidade de um motor onde alunos e professores interajam, direcionando resultados certos para a busca pretendida. Aponta como erro o retorno de publicidades e conteúdos indesejados e de forma incompreensível para o pesquisador. Levanta ainda o problema com as telas, que muitas vezes não adotam o perfil funcional em celulares e tablets, que são usados pelo público.

Semelhante a essa ideia, Sandra Regina Rocha Silva e Geraldo Bonorino Xexéo (2013), propõem uma forma colaborativa, onde o público infantil tenha resultados de buscas filtrados pelos pais. Percebesse assim, que as tarefas do motor de busca não estão otimizadas o suficiente, a ponto de autonomia para deliberar o que é um resultado relevante ou não.

Autores como Chuen-Tsai Sun, Shu-Hao Ye e Hsi-Chien Hsieh (2017), trazem a ideia onde devesse ter melhoramento no algoritmo de busca, levando em conta característica; de interface, facilidade de compreensão. Apesar de existirem motores de buscas preocupados com este item, acabam não cumprindo os demais requisitos para atender esta demanda.

Dania Bilal e Meredith Boehm (2013) propõem que os ranques existentes não levam a questão da idade com a relevância necessária, de modo a colocar o que importa e o que é esperado para os alunos, não causando assim a frustração por parte do pesquisador. Esta preocupação é de conhecimento geral, e alinhado com as ideias anteriores, nos remete que crianças têm grande possibilidade de acessar resultados errados em sua pesquisa.

7. CONCLUSÃO

Os trabalhos apresentam, de modo geral, a importância da qualificação no momento de realizar busca em um motor de pesquisa [Hahnel et al., 2018; Faslah, 2016; Sun, Ye e Hsieh, 2017]. Aponta ainda que é possível realizar uma otimização no processo e que se pode treinar para isso [Faslah, 2016]. Todos os cenários observados acontecem no meio acadêmico, em sua maioria com alunos secundários, que se aproxima com o objeto central da pesquisa aqui proposta. Destaca-se ainda o baixo índice de pesquisa envolvendo a temática de mecanismo de busca aplicado a educação básica, apesar de o acesso à internet por crianças ser crescente e funcionar como um mecanismo de auxílio na aprendizagem ativa.

Foram de grande importância os títulos escolhidos, pois tornou-se possível identificar alguns padrões entre os trabalhos, como a existência de deficiência no momento de busca de alunos na Web, a falta de um ambiente propício para tal [Azpiazu et al., 2017]. Observou-se ainda a existência de motores de busca que buscam traçar o perfil do usuário, ou ainda de ter uma busca colaborativa, de modo a atender o perfil das crianças [SILVA, XEXÉO, 2013; VERMA et al. 2014].

Os trabalhos analisados permitem ainda inferir os seguintes aspectos relevantes a serem considerados, de acordo com a literatura aqui apresentada:

1. Filtrar conteúdos impróprios;
2. Possuir visual amigável;
3. Gerar recompensas satisfatórias;
4. Possibilitar o acesso através de várias plataformas (desktop, celular, etc.);

5. Classificar da busca por algum especialista; e
6. Sugerir de novas buscas, seguindo o padrão de uma anterior.

Conclui-se assim que a temática é de relevante e pode ser trabalhada em trabalhos futuros. Notasse ainda que a pergunta de revisão é respondida, através de demonstrações e interesse no tema, notasse que existem ambientes inovadores envolvendo a temática, porém que ainda não estão conseguindo gerar o impacto desejado, possibilitando extrair aspectos de relevância para o tema.

Referencias

- Azpiazu, I. M., et al (2017) *Online Searching and Learning: YUM and Other Search Tools for Children and Teachers*. Information Retrieval Journal, 20.5: 524-545.
- Babayigit, S. e Groppe, S. (2018) *Webpage Ranking Analysis of Various Search Engines With Special Focus on Country-Specific Search*. Open Journal of Web Technologies (Ojw), V. 5, N. 1, P. 44-64.
- Bilal, D. e Boehm, M. (2017) *Towards New Methodologies for Assessing Relevance of Information Retrieval from Web Search Engines on Children's Queries*. Qualitative and Quantitative Methods in Libraries, 2.1: 93-100.
- Baldin, N. e Munhoz, E. (2011) Snowall (Bola De Neve): Uma Técnica Metodológica para Pesquisa em Educação Ambiental Comunitária. X Congresso Nacional De Educação – Educere.
- Faslah, R. (2016) *Model of The Use of Information Technology in Improving the Ability of Student*. Jurnal Ilmiah Econosains, 14.1: 67-81.
- Gossen, T., Kotzyba, M. e Nürnberger, A. (2017) *Search Engine for Children: User-Centered Design*. Datenbank-Spektrum, 17.1: 61-67.
- Hahnel, C., et al. (2018) *The Role of Reading Skills in the Evaluation of Online Information Gathered from Search Engine Environments*. Computers in Human Behavior, 78: 223-234.
- Internet Society. (2017). “Internet Society Global Internet Report 2017”. <https://Future.Internetsociety.Org/Wp-Content/Uploads/2017/09/2017-Internet-Society-Global-Internet-Report-Paths-To-Our-Digital-Future.Pdf>, Dezembro.
- Pera, M. S. e Ng, Y. (2017) *Using Online Data Sources To Make Query Suggestions For Children*. In: Web Intelligence. Ios Press. P. 303-323.
- Silva, S. R. R. e Xexéo, G. B. (2013) *Child Search Framework: A Collaborative Information Retrieval Architecture to Assist Children in the Search Process*. In: Computer Supported Cooperative Work in Design (Cscwd), Ieee 17th International Conference On. Ieee, P. 551-556.
- Sun, C., Ye, S. e Hsieh, H. (2014) *Effects of Student Characteristics and Question Design on Internet Search Results Usage in a Taiwanese Classroom*. Computers & Education, 77: 134-144.
- Vanderschantz, N. e Hinze, A. (2017) *How Kids See Search: A Visual Analysis of Internet Search Engines*. In: Proceedings of The 31st British Computer Society Human Computer Interaction Conference. Bcs Learning & Development Ltd.. P. 84.
- Verma, D., Minocha, K. e Kochar, B. (2014) *A Multi-Agent Based Personalized Search Engine with Topical Crawling Capabilities*. Iup Journal of Computer Sciences, 8.3: 20.