

Um mecanismo de busca para apoio a aprendizagem infantil baseado em mineração de dados

Laercio Pontin Jr., Heleno Fülber, Adonney Allan O. V., Bruno Merlin

Programa de Pós-Graduação em Computação Aplicada – Universidade Federal do Pará
(UFPA)
68464-000 – Tucuruí – PA – Brasil

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

Abstract. *The technology can still be improved when used in teaching. One example is web search tools for elementary education, which can be a complex environment and disadvantageous to learning. This is due to the little familiarity of children in formulating questions and the difficulty in interpreting the results from the search engine. Therefore, this work proposes a search engine that improves the search string considering what governs the Base Nacional Comum Curricular. To implement this mechanism, a text mining technique is used, to improve the environment for children's learning. In this way, it is hoped to improve the results of web search aimed at learning by the public.*

Resumo. *A tecnologia ainda pode ser melhorada quando utilizada no ensino. Um exemplo são ferramentas de pesquisa na web para ensino fundamental, que pode ser um ambiente complexo e que desfavoreça a aprendizagem. Isso ocorre devido a pouca familiaridade de crianças na formulação de perguntas e a dificuldade em interpretar os resultados provindos no motor de busca. Logo, este resumo 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 técnica de mineração de texto, de forma a tornar o ambiente aprimorado para aprendizagem infantil. Espera-se assim melhorar os resultados de busca web voltado a aprendizagem do público.*

1. Introdução

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, e 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 e a sumarização dessas informações para disponibilizar uma maneira fácil e rápida de ajudar os usuários a encontrá-las, denominamos esses sistemas como motores de busca.

Neste universo virtual, denominado internet, possui uma infinidade de sites, com as mais diversas funcionalidades e informações. Logo, para conseguir encontrar alguma informação dentro da *web* surgiram os 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 usuários dos motores pesquisem qualquer tipo de informação através de palavras-chave, Costa (2011).

Segundo Aresta, et al. (2008) a adesão à utilização dessas ferramentas poderá 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.

No entanto, quando se pensa em modelo de aprendizagem infantil no ensino fundamental, os motores de pesquisa difundidos na internet não possuem uma eficiência. 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 o nível de leitura avançada sobre temas complexos que as crianças têm dificuldade de compreender [PERA e NG, 2017].

Desta forma, esta pesquisa busca elaborar um mecanismo de busca que aprimore os resultados de busca voltados a aprendizagem de aluno do ensino fundamental. Esse mecanismo tem como diferencial uma abordagem não supervisionada, de forma que a avaliação dos melhores resultados para a pesquisa será feita por intermédio da Base Nacional Comum Curricular -BNCC (documento oficial que norteia a educação básica). É utilizado ainda técnicas de mineração de texto, que permite relacionar o assunto buscado pela criança com o que é regido na BNCC.

2. Desenvolvimento

Neste capítulo é realizado o desenvolvimento da pesquisa. Para tal foram definidas as tecnologias, criadas as etapas pertencentes a mineração de texto, conforme Figura 1. 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 portal *web*, que permite ao usuário ter a experiência com a ferramenta, tendo um melhor retorno na pesquisa *web*, de acordo com seu perfil.



Figura 1. Fluxograma de trabalho de mineração, criada pelo autor.

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).

Na análise e extração do conhecimento foi o momento em que se realizou a seleção dos dados, momento ao qual se retirou as informações da BNCC e na sequência

foram inseridas em *arrays*, que permitem seguir para o processo de preparação dos dados. Nesta preparação ocorrem as técnicas de tokenização nas informações, cujo objetivo é reduzir a numerosidade de palavras repetidas ou desnecessárias do texto (espaço vazio, caracteres especiais, etc).

Ocorre ainda a retirada das *stopwords* (que são os termos que nada acrescentam à representatividade da coleção ou que sozinhas nada significam, como artigos, pronomes e advérbios), desta forma ficam no texto apenas palavras que agregam no momento da busca na *web*. O último processo feito na extração é conhecido como *Stemming* cujo objetivo é transformar as palavras em seu radical, de forma a diminuir a numerosidade. Com essas técnicas, a quantidade de palavras que inicialmente eram de 276045 foi reduzida para 25297, a serem armazenadas no banco de dados.

Após a redução de numerosidade, foi possível a criação da base de dados que atua no processo de pós-processamento, para tal criou-se a base de dados seguindo o modelo abaixo. Nesta base foram inseridos os dados extraídos da BNCC.

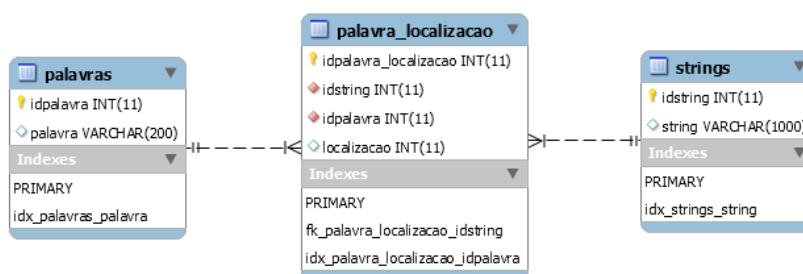


Figure 2. DER banco de dados do mecanismo de busca, criada pelo autor.

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 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.

Uma vez implementado 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. Sua implementação seguiu padrões semelhantes ao do Google, com o visual da Figura 3.

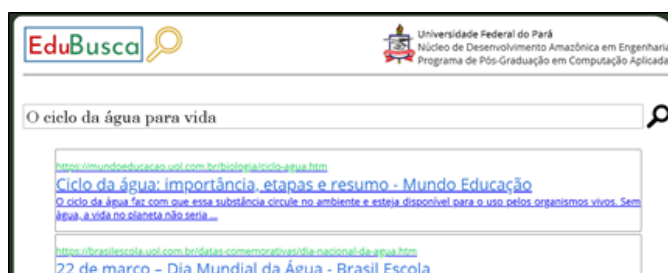


Figure 3. Visual do mecanismo de busca, criada pelo autor.

3. Resultados e discussões

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.

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 demonstram 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 portal *web*. O processo de classificação de busca de forma positiva termos correlatos.

Assim, conclui-se que este projeto atingiu seu objetivo inicial de forma satisfatória, gerando 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.

Referências

- Aresta, M; Moreira, A. e Pedro, L. (2008) “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.
- Costa, F. M. S. P. Search (2011) “Engine and Optimization.” 63f. Dissertação (Mestrado em Sistemas e Tecnologias de Informação para as Organizações) apresentada ao Instituto Politécnico de Viseu/POR.
- 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.
- Xavier, B.; Silva, A.; Gomes, G. (2015) “Uma arquitetura híbrida para a indexação de documentos do Diário Oficial do Município de Cachoeiro de Itapemirim.” *Transinformação*, v. 27, n. 1, p. 83-95.