FORMULÁRIO PARA RELATÓRIO FINAL

1. Identificação do Projeto

Título do Projeto PIBIC/PAIC

UM ESTUDO SOBRE O IMPACTO DA FUNÇÃO DE SIMILARIDADE SOBRE ALGORITMOS DE PROCESSAMENTO DE CONSULTAS EM SISTEMAS DE BUSCA TEXTUAL.

Orientador

Prof. Dr. Edleno da Silva Moura

Aluno

Lucas Citolin

2. Informações de Acesso ao Documento

2.1 Este documento é confidencial?

[ ] **SIM** [x] **NÃO**

2.2 Este trabalho ocasionará registro de patente?

[ ] **SIM** [x] **NÃO**

2.3 Este trabalho pode ser liberado para reprodução?

[x] **SIM** [ ] **NÃO**

2. 4 Em caso de liberação parcial, quais dados podem ser liberados? Especifique.

Qualquer dado relacionado a pesquisa é público.

3. Introdução

Sistemas de busca são complexas ferramentas utilizadas pelo usuário com o propósito de procurar informação em um conjunto de documentos sem nenhuma semelhança inicial. A recuperação de informação em sistemas de busca textual ocorre em basicamente três etapas: Operação de consulta, na qual o usuário entra com os termos para a pesquisa, sendo estes termos na maioria dos casos palavras-chaves. Indexação**,** na qual ocorre a criação das estruturas de dados que serão, posteriormente, utilizadas durante o processamento, sendo uma estrutura muito utilizada o índice invertido. Estas estruturas serão preenchidas pelos termos dos documentos e suas respectivas importâncias para o documento, a qual há de ser determinada através da função de pesos adotada. Finalmente, o processamento de consultas, na qual ocorre a pesquisa no índice criado durante a indexação. Os documentos julgados como importantes para consulta são por fim retornados ao usuário, ordenados pela sua importância.

Nota-se que, apesar de ser um processo sistemático, existem duas variáveis no processo de busca: Primeiro, a função matemática que irá calcular a importância de cada termo para seus respectivos documentos, conhecida como função de peso ou similaridade. Segundo, o algoritmo de processamento de consultas que irá escolher de que maneira tratar os dados indexados durante o cálculo da similaridade.

Este projeto visa o estudo do impacto das funções de similaridade sobre algoritmos de processamento de consultas em sistemas de busca textual. Apesar da importância desses dois fatores na implementação de sistemas de busca, surpreendentemente há uma lacuna na literatura quando busca-se trabalhos que estudem a relação entre a função de similaridade e o desempenho dos principais algoritmos de processamento de consultas propostos até hoje. O objetivo é comparar o desempenho dos principais algoritmos de processamento de consultas existentes na literatura quando aplicados com diferentes funções de similaridade.

4. Justificativa

A utilização de sistemas de busca cresce de acordo com a quantidade de dados armazenados. Quando os dados são grandes coleções de documentos, exige-se uma efetiva busca por meio de palavras chaves fornecidas pelo usuário. O crescimento da coleção de documentos e a falta de acompanhamento do potencial de processamento das máquinas atuais levou ao investimento na área conhecida como recuperação de informação.

O constante investimento no processamento de consultas tem como objetivo o retorno de algoritmos efetivos (velozes, precisos) que acompanhem o grande crescimento de dados armazenados. Para que a busca seja efetuada na coleção de documentos, utiliza-se métodos para o cálculo de similaridade entre dois documentos, sendo este o fator decisivo na velocidade e precisão do (algoritmo) processo de busca.

Fórmulas para o cálculo de similaridade se tornam presente na literatura desde modelos mais tradicionais como o Modelo de Espaço Vetorial e o modelo BM25. Existe uma grande lacuna na implementação destes modelos: a falta de comparação entre eles. Uma função de similaridade pode ser implementada em algoritmos diferentes, fazendo parte essencial do desempenho do mesmo.

Os trabalhos atuais não fazem uma efetiva comparação de seus algoritmos com modelos de similaridade diferentes, deixando uma lacuna na literatura pela combinação de modelos mais efetiva. A proposta deste projeto é de implementar os algoritmos de processamento de consultas, tais como o WAND, BMW e o CSP com diferentes modelos para o cálculo de similaridade.

5. Objetivos

Quando um novo algoritmo é proposto pela comunidade de RI a literatura, em muitas das vezes o autor do artigo não submete o seu algoritmo a testes com diferentes funções de similaridade, limitando seu ambiente de testes e não abrangendo todas as possibilidades.

Este artigo tem como objetivo provar empiricamente que o desempenho dos algoritmos de busca textual varia de acordo com a função de similaridade aplicada. O intuito não é avaliar a qualidade do processamento de busca quando aplicado com diferentes funções de peso, mas sim o desempenho de tempo de execução do algoritmo, justificando que os algoritmos propostos sejam submetidos a diferentes implementações com as funções de peso para que seu desempenho seja bem conhecido.

6. Metodologia

Primeiro, fora realizado um estudo aprofundado da literatura com o objetivo de obter conhecimento sobre os principais trabalhos relacionados, fixar conceitos básicos necessários para a realização do trabalho e também dar embasamento para a fase de implementação de métodos e algoritmos a serem desenvolvidos durante o projeto.

Posteriormente, ocorreu a implementação de arcabouços de programação necessários para o desenvolvimento do projeto. Após o desenvolvimento do arcabouço de programação, ocorreu exaustivos experimentos, assim como a análise dos mesmos e seus significados para a literatura.

Uma parte fundamental do trabalho é o planejamento de experimentos. Em boa parte dos experimentos realizados no projeto são manipuladas grandes bases de dados de informação. A validação dos métodos estudados será feita por meia de experimentação exaustiva utilizando-se técnicas tradicionais utilizadas na área, tais como avaliação de eficiência em termos de tempo e recursos computacionais exigidos pelos sistemas, avaliação de curvas de precisão e revocação, desenvolvimento de testes de significância dos resultados obtidos, tais como t-test para a comparação entre sistemas de busca.

7. Resultados e Discussão

Os testes foram realizados combinando os algoritmos BMW, WAND e CSP com as diferentes funções de peso citados neste artigo, já conhecidas pela literatura: BM25, Vetorial e Vetorial sem a utilização da norma. Cada dado apresentado nestes resultados é uma média de quatro testes realizados, visto que existe uma pequena oscilação de tempo a cada execução.

Os algoritmos BMW e WAND não dispõe da divisão do índice em camadas durante a indexação, diferente do algoritmo CSP, o qual busca dividir o índice em dois com o intuito de otimizar a busca. Pelos algoritmos apresentados neste artigo trabalharem com indexação documento-a-documento, o indíce é organizado de tal maneira que as listas mais robustas do índice invertido (os documentos com mais termos) estejam no começo do índice, pois serão os mais acessados durante a busca. Tirando proveito desta caracteristica, o algoritmo CSP tende a dividir o índice em duas parcelas, processando os dois indices sequencialmente, assim podendo encontrar o documento a ser buscado logo no primeiro índice sem precisar buscar no segundo, funcionando como um mecanismo de poda.

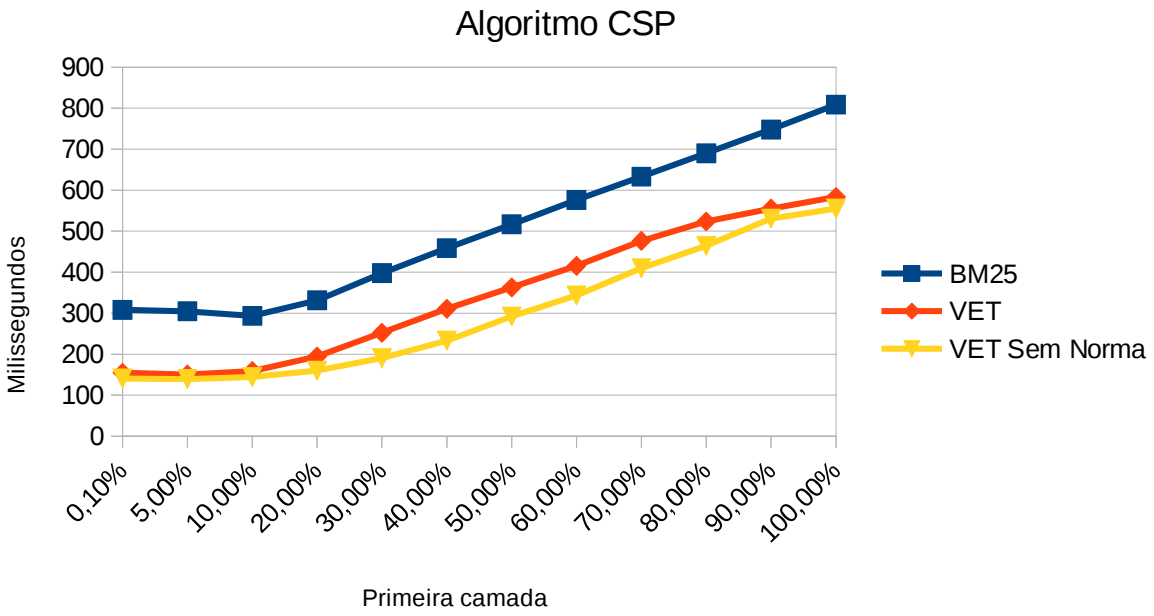

Figura 1: Algoritmo CSP quando aplicado as diferentes funções de similaridade em diferentes distribuilção de camadas

A Figura 1 apresenta os resultados dos testes do processamento do algoritmo CSP quando aplicado com diferentes funções de peso: BM25, Vetorial e Vetorial sem norma e sua distribuição em camadas, sendo o eixo das abscissas o tamanho da primeira camada do índice. Podemos visualizar pelo gráfico a melhor divisão em camadas para cada função de similaridade quando aplicada no algoritmo CSP. Para a função Vetorial e Vetorial sem norma, encontramos sua melhor distribuição em 5% do índice na primeira camada, já para o BM25, seu melhor ponto pode ser identificado quando o primeiro índice contém 10% do total.

Este experimento fora crucial para confirmar que as funções de similaridade não afetam apenas o desempenho de tempo final dos diferentes algoritmos, mas também na distribuição em camadas que deve ser usada durante a indexação.

A comparação entre os resultados do método WAND, BMW e CSP (com seus respectivos melhores pontos) pode ser encontrado na Figura 2. Neste gráfico podemos visualizar não apenas a evolução dos algoritmos que foram apresentados em ordem cronológica, mas o impacto que as diferentes funções de peso tem sobre estes.

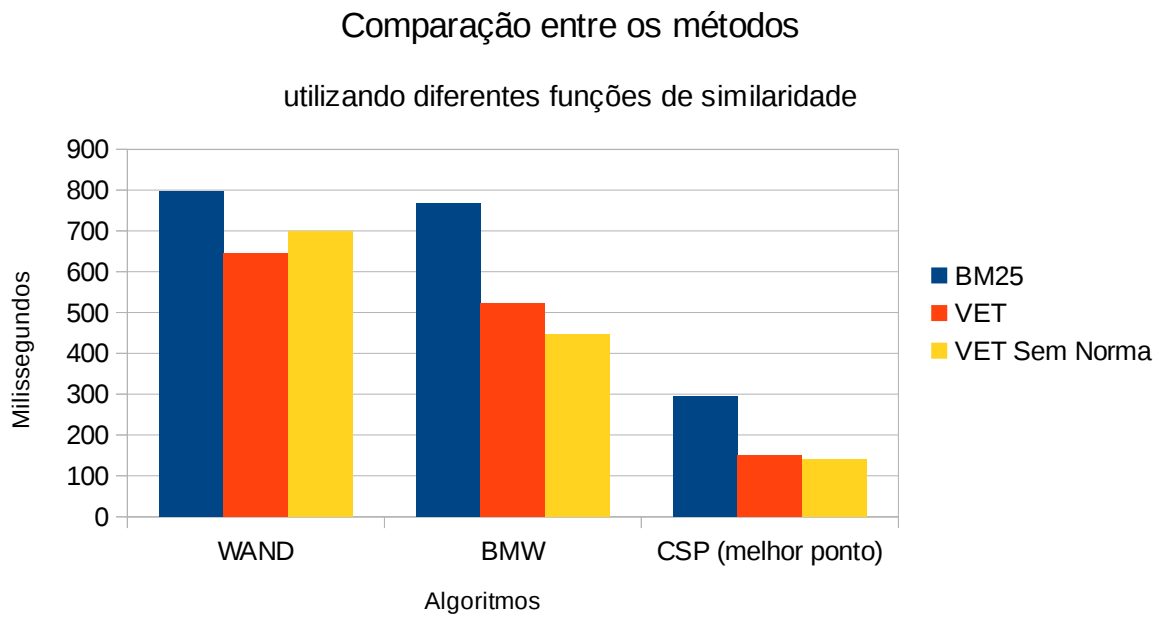

Figura 2: Resultado dos algoritmos WAND, BMW e CSP quando aplicados com as diferentes funções de similaridade BM25, Vetorial e Vetorial sem a norma.

Com este gráfico da Figura 2 fica evidente que a utilização de diferentes funções de similaridade altera o tempo de execução do algoritmo de maneira não sistemática. O algoritmo WAND apresenta melhor desempenho com o modelo Vetorial, já o BMW apresenta seu melhor desempenho quando aplicado com o modelo Vetorial sem a norma. Podemos concluir que não existe uma função de similaridade melhor, ou que apresente um melhor desempenho geral entre os algoritmos, mas sim funções de similaridade que, dependendo do algoritmo implementado, terá seu desempenho especifico.

Esta análise é fundamental a literatura pois deixa evidente que os algoritmos já existentes e os futuros a serem propostos devem passar por uma verdadeira combinação com as diferentes funções de similaridade para que se possa afirmar seu melhor desempenho.

Conclui-se então que o desempenho de tempo de execução de um algoritmo de busca textual não pode ser medido sistemáticamente como vem sendo feito pela literatura. Esta pesquisa demonstrou empiricamente, atráves de diversos testes realizados entre as possíveis combinações de algoritmo-função de peso, que um algoritmo apresenta significativamente diferentes tempos de execução quando implementado com diferentes cálculos de similaridade.

Para algoritmos que desfrutam de índice dividido em camadas, o impacto desta pesquisa é ainda maior: o algoritmo pode apresentar tempos radicalmente diferentes,

como analisado na Figura 1: onde o desempenho entre a implementação do método CSP com Vetorial e Vetorial sem norma demonstra-se irregular, com seu tempo de execução dependendo muito da divisão em camadas implementada.

Os testes propostos e realizados neste artigo tiveram como objetivo medir o desempenho de tempo de execução do WAND, BMW e CSP com as funções de peso apresentadas. Porém, o termo desempenho abrange diferentes categorias não analisadas neste projeto, como a similaridade dos documentos retornados com a consulta em cada combinação de algoritmo-função de similaridade, abrindo assim a possibilidade para trabalhos futuros.

8. Referências

Robertson, S. E. & Walker, S. (1994). **Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval**. Em ACM SIGIR, pp. 232--241.

Salton, G.; Wong, A. & Yang, C. S. (1974). **A vector space model for automatic indexing**. Relatório técnico, Ithaca, NY, USA.

Moura, E. S. D.; Fernandes, D. R.; Silva, A. S.; Calado, P. & Nascimento, M. A. (2005). **Improving Web Search Efficiency via a Locality Based Static Pruning Method.** World Wide Web.

Rossi, Cristian; de Moura, Edleno S.; CARVALHO, ANDRE L.; da Silva, Altigran S. **Fast document-at-a-time query processing using two-tier indexes**. In: the 36th international ACM SIGIR conference, 2013, Dublin. Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval - SIGIR '13. New York: ACM Press, 2013. p. 183-192.

Ding, S. & Suel, T. (2011). Faster top-k document retrieval using block-max indexes. Em ACM SIGIR, pp. 993--1002.

9. Cronograma de Atividades

Nº	Descrição	Ago 2015	Set	Out	Nov	Dez	Jan 2016	Fev	Mar	Abr	Mai	Jun	Jul
1	Revisão da Literatura	X	X										
2	Implementação dos métodos e funções de similaridade		X	X	X	X							
3	Elaboração de Relatório parcial				X	X	X						
4	Experimentos com funções de similaridade					X	X	X	X	X	X		
5	Busca de alernativas para aprimorar algoritmos existentes e adapta-los a funções de similaridade estudada							X	X	X	X		
6													
7													
8													
9	- Elaboração do Resumo e Relatório Final (atividade obrigatória) - Preparação da Apresentação Final para o Congresso (atividade obrigatória)											X	X