Máquina de busca - jsearch

Josué Santos Silva

Resumo

O relatório apresenta informações sobre a implementação algoritmos básicos para a execução do trabalho assim como tecnologias e ferramentas utilizadas no processo. Na sequência é apresentado a proposta do módulo Expansor de Consultas para a máquina de busca. Na conclusão serão apresentados os resultados comparativos de acordo com as métricas estabelecidas entre os algoritmos básicos e o módulo proposto.

Palavras-chaves: recuperação da informação. relevance feedback. rocchio. bm25.

Introdução

O projeto está separado logicamente em componentes Indexador, Componente Ranqueador, Componente Expansor de Consultas, Componente Gerador de Log. Ao final aspectos de efetividade (qualidade do ranking gerado em MAP, P@5 e nDCG

Desenvolvimento

- Ferramentas
  - Lucene
  - ObjectDB
  - Java 8
  - HTML/JS

Componente Indexador

- Compactação
  - LZ4

- Foi realizada uma modificação no Analyzer do Lucene, para que tanto a query inserida pelo usuário, quanto os algoritmos de ranking considerem a remoção de stopwords(este projeto considera apenas idioma inglês), tokenização e stemming(PorterStem)

Componente Ranqueador

- O ranking da máquina de busca proposta utiliza a função de ranking BM25(Parâmetros k1 = 1.2 e b = 0.75).

- O componente de ranking também possui um mecanismo de expansão de de consulta, no qual foi implementado o algoritmo Rocchio(Parâmetros alpha = 1.0 e beta = 0.8)

Componente Expansor de Consultas

- O componente expansor foi implementado baseado no Log de consultas e em uma varredura prévia pelos documentos indexados, em busca de termos importantes.

- Na primeira parte são verificados no log consultas, consultas semelhantes previamente efetuadas e seus respectivos documentos relevantes para os usuários.

Componente Expansor de Consultas

- Na segunda parte é efetuada uma busca com o BM25, e são retornados os 10 melhores resultados. A partir deles são retirados termos, que possuem um peso.

- Uma lista de termos ordenada por tf*idf, além de serem atribuídos pesos em um fator de 1.2 para termos que ocorrem no título dos documentos.

Componente Expansor de Consultas

- Na segunda parte é efetuada uma busca com o BM25, e são retornados os 10 melhores resultados. A partir deles são retirados termos, que possuem um peso. Por fim é gerada uma lista de termos ordenada por peso.

- Por fim os 10 melhores termos são concatenados a consulta original.

Componente Gerador de Log

- O componente gerador de Log utiliza o banco de dados ObjectDB, um sistema de banco de dados relacional para armazenamento de objetos.

- Para cada consulta são armazenados IP de origem do usuário que efetua a consulta, a consulta, o id do documento no qual o usuário interagiu e o timestamp em que a consulta foi realizada.

Componente Gerador de Log

- O recurso de autocompletar recupera as consultas realizadas pelo respectivo usuário considerando a princípio as consultas em um intervalo de 24 horas. São retornados no máximo 5 resultados.

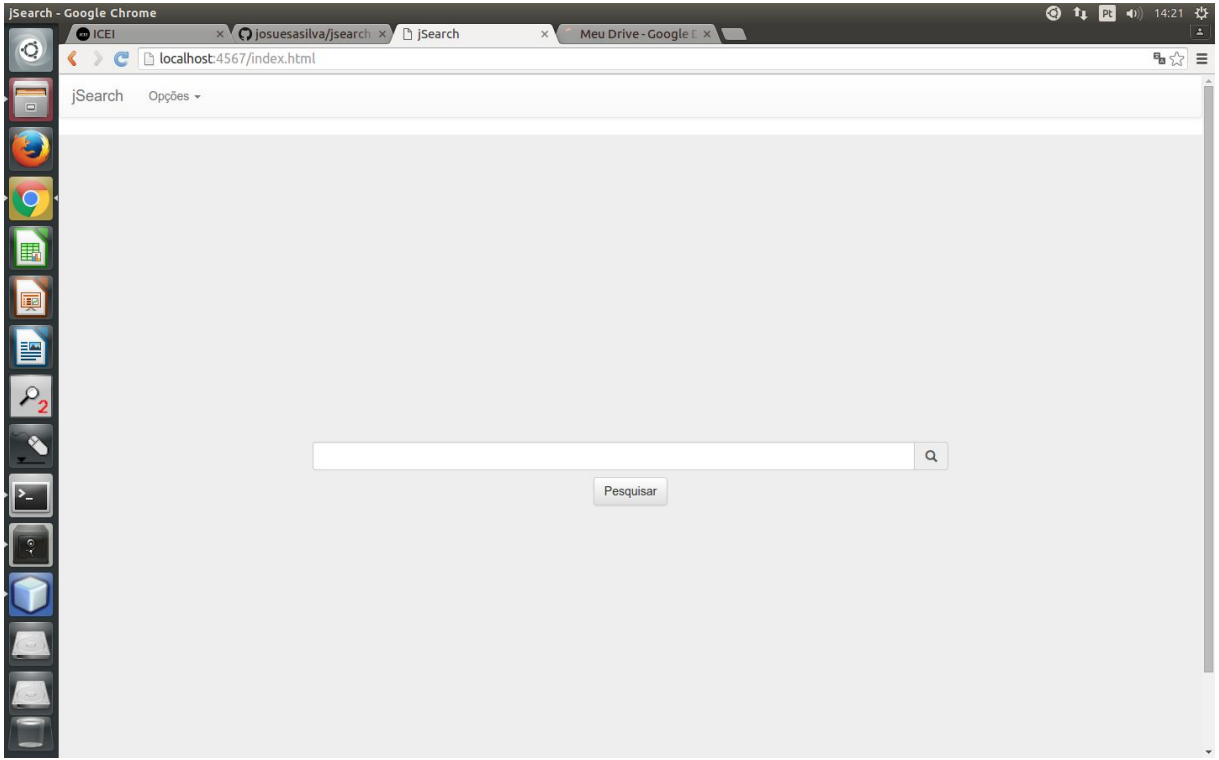
- Para autocompletar, o componente verifica a similaridade entre a consulta original do usuário e as demais consultas realizadas pelo usuário contidas na base de dados e retorna as cinco melhores.

Componente Gerador de Log

- O critério de similaridade utilizado é baseado na Distância de Jaro-Winkler.

- O algoritmo retorna um fator, que quanto maior mais semelhante são os strings comparados. Foi utilizado como fator de corte o valor de 0.8, ou seja, 80% de semelhança entre as consultas comparadas.

Interface Gráfica



Resultados

Os testes foram realizados em um notebook Dell Vostro 5470, com processador i5(dois núcleos e 4 threads) quarta geração, 4Gb de memória ram DDR3, armazenamento SSD.

Desempenho do indexador

Documentos indexados	17123	30879	1697253
Tempo de indexação	20,31 seg.	65,411 seg.	2340 seg.
Documentos indexados por segundo	843	472	725
Tempo de resposta a consultas ao índice	0,07 seg.	0,08 seg.	0.052
Taxa de compressão	25%	25%	25%

Desempenho do ranking (por consulta)

Algoritmo	Tempo de resposta
BM25	10 ms
Rocchio	47 ms
Expansor de consultas	85 ms

Desempenho do gerador de log

Operação	Tempo de resposta
Armazenar consulta	344 ms
Recuperar log	78 ms

Resultados

Em termos gerais, partindo do oráculo da coleção do TREC e as respectivas queries, não foram obtidas melhoras em relação ao BM25. Isso ocorre pelo fato de que o expansor de consultas proposto levar em consideração interações prévias dos usuários, ou seja, a partir do log de consultas são avaliados documentos relevantes e deles são extraídos termos mais relevantes para o usuário, ou partindo de algo que foi colocado no log para viabilizar consultas mais eficientes.

Resultados

Ainda de acordo com os testes efetuados partido do oráculo e das queries do TREC, avaliando cada consulta separadamente, foi notado que algumas poucas consultas obtiveram melhores resultados com o expansor de consultas. Consultas muito pequenas, em alguns momentos, foram beneficiadas, já em outros momentos nem tanto, pelo fato de se tornar ambígua. Consultas maiores (mais que 5 tokens) foram mais beneficiadas que consultas pequenas.

Resultados

Consta no repositório no Github a planilha com os resultados das demais métricas requisitadas. Além de mais detalhes no Relatório Técnico.