Desempenho de Busca: Whoosh vs Elasticsearch

Carlos Eduardo Alves Ferreira

Universidade Federal do Rural do Rio de Janeiro (UFRRJ)
CEP 26285-060 – Rio de Janeiro – RJ – Brazil

alvesgt3031@uffrj.br

Abstract. *The article compares the Whoosh and Elasticsearch search engines using graphs and data. Elasticsearch demonstrates significant superiority in speed, being 14 times faster in indexing (1.02s vs. 14.86s) and 7 times faster in search (1.40s vs. 10.14s) compared to Whoosh. Additionally, Elasticsearch shows a slight improvement in retrieving relevant results, especially in the 1 x k search. In conclusion, for applications involving the indexing and searching of large datasets, Elasticsearch proves to be a faster and more effective option than Whoosh.*

Resumo. *O artigo compara os motores de busca Whoosh e Elasticsearch usando gráficos e dados. O Elasticsearch demonstra superioridade significativa em velocidade, com 14 vezes mais rapidez na indexação (1,02s vs. 14,86s) e 7 vezes mais velocidade na busca (1,40s vs. 10,14s) em comparação com o Whoosh. Além disso, o Elasticsearch apresenta ligeira melhoria na recuperação de resultados relevantes, especialmente na busca 1 x k. Em conclusão, para aplicações que envolvem a indexação e busca de grandes conjuntos de dados, o Elasticsearch é uma opção mais rápida e eficaz do que o Whoosh.*

1. Introdução

A busca e recuperação de informação (BRI) é uma área da ciência da computação que se concentra no desenvolvimento de sistemas que permitem aos usuários encontrar informações relevantes em grandes conjuntos de dados. Os motores de busca são uma das principais tecnologias utilizadas para BRI.

Existem muitos motores de busca disponíveis, cada um com suas próprias vantagens e desvantagens. Neste artigo, comparamos dois motores de busca populares, o Whoosh e o Elasticsearch. O Whoosh é uma biblioteca de pesquisa e indexação de texto completo rápida, programável e fácil de usar. O Elasticsearch é um mecanismo de pesquisa e análise de dados RESTful distribuído.

Avaliamos os dois motores de busca com base em três métricas: tempo de indexação, tempo de busca e precisão/recall.

2. Analisando o conjunto de dados

O conjunto de dados Cranfield é uma coleção clássica de dados usada em Busca e

Recuperação de Informação (BRI) e tem sido uma referência na área desde a década de 1960. A análise do conjunto de dados Cranfield envolve a compreensão de seus componentes principais: os resumos científicos (cran.all.1400), as consultas (cranqrel), e os resultados relevantes para as consultas (cranqr).

O arquivo cran.all.1400 abriga os resumos dos documentos, cada um identificado por um número único. Cada entrada de documento inclui informações como título, autor, bibliografia e corpo do texto. Em paralelo, o arquivo cranqry armazena as consultas, cada uma atribuída a um número de identificação e acompanhada de seu texto correspondente. Por fim, o arquivo cranqrel contém as avaliações de relevância, fornecendo, para cada consulta, uma lista dos documentos considerados relevantes.

Juntos, esses três arquivos formam um conjunto de dados que pode ser usado para testar algoritmos de Recuperação de Informação. Você pode usar as consultas no arquivo cranqry para buscar documentos no arquivo cran.all.1400 e, em seguida, comparar seus resultados com as avaliações de relevância no arquivo cranqrel. Os componentes do conjunto de dados permitem a avaliação de sistemas de BRI em diferentes aspectos, incluindo a capacidade de indexação, a qualidade das consultas e a eficácia na recuperação de documentos relevantes.

3. Whoosh vs Elasticsearch: Comparando Ferramentas para Busca de Texto

Elasticsearch e Whoosh são duas ferramentas poderosas usadas para pesquisar e recuperar informações. Ambos são projetados para lidar com grandes volumes de dados, mas cada um tem seus próprios recursos e benefícios.

Elasticsearch é uma ferramenta de código aberto criada para pesquisar e processar grandes quantidades de dados e opera como um banco de dados não relacional1. Foi desenvolvido por Shay Bannon e lançado pela primeira vez em 2010. Baseado em tecnologias como Apache Lucene e Java, o Elasticsearch possui uma interface simples e estruturada que suporta HTTP e JSON. Para ter a elasticidade necessária, os recursos têm melhor desempenho quando distribuídos em clusters1. Portanto, pode fornecer a escalabilidade necessária para trabalhar com ferramentas de análise de dados baseadas em big data. Elasticsearch é um mecanismo distribuído de pesquisa e análise de dados RESTful capaz de atender a um número crescente de casos de uso. No coração do Elastic Stack, ele armazena centralmente seus dados para fornecer pesquisa rápida, relevância refinada e análises poderosas e facilmente escaláveis.

Whoosh é uma biblioteca de pesquisa e indexação de texto completo rápida, programável e fácil de usar. É Python puro, o que significa que você não precisa compilar ou instalar nenhuma extensão binária para usá-lo. Whoosh permite adicionar funcionalidade de pesquisa de texto completo ao seu aplicativo ou site Python. Ele permite pesquisar texto completo, frases, prefixos e intervalos, ao mesmo tempo que tem controle total sobre o esquema de indexação. Ele também oferece suporte a pesquisas rápidas de texto, mesmo com grandes volumes de texto.

Ambas as ferramentas são ótimas opções para sistemas de busca e recuperação de informações, dependendo das necessidades específicas do seu projeto.

4. Avaliação de sistemas de recuperação de informação

A avaliação de sistemas de busca e recuperação de informação é um componente fundamental no desenvolvimento e aprimoramento dessas plataformas, visando garantir eficácia e relevância na entrega de resultados aos usuários. Neste contexto, diversas métricas são empregadas para mensurar a performance, destacando-se precision, recall, tempo de busca e tempo de indexação.

4.1. Avaliação do tempo de indexação

No centro dos sistemas de recuperação de informação, o tempo de indexação se destaca como o peça fundamental responsavél pela eficiência do sistema. Esta métrica, que define o tempo necessário para criar ou atualizar o índice de busca, tem um papel vital na determinação da rapidez e eficácia do processo de busca.

O tempo de indexação não é apenas um indicador técnico; é uma estratégia essencial para otimizar o desempenho geral do sistema. Uma abordagem eficaz neste aspecto tem um impacto direto na velocidade com que o sistema pode fornecer informações relevantes aos usuários, melhorando positivamente a experiência de busca.

Quando o tempo de indexação é gerenciado de forma eficiente, ele ajuda a acelerar a resposta do sistema às consultas dos usuários. Uma indexação rápida e precisa implica que as atualizações ou adições de novos dados são integradas ao índice de forma eficiente, refletindo diretamente na rapidez do processo de busca.

O tempo de indexação e o tempo de busca são componentes interconectados. Uma indexação eficiente facilita uma busca mais rápida e responsiva, criando uma sinergia que contribui para uma experiência de usuário integrada e satisfatória.

4.2. Avaliação do tempo de busca

O tempo de busca é um elemento fundamental na avaliação de sistemas de recuperação de informação, desempenhando um papel significativo na formação da experiência do usuário. A rapidez é fundamental para fornecer respostas prontas e eficazes, correspondendo às expectativas atuais dos usuários que desejam um acesso rápido e eficaz à informação.

A magnitude do conjunto de dados e a complexidade das consultas são fatores cruciais que influenciam diretamente o tempo de busca. Grandes conjuntos de dados e consultas complexas podem testar a eficiência do sistema, tornando essencial a adoção de estratégias eficientes para otimizar esse processo.

A eficácia na construção do índice de busca é um elemento essencial para reduzir o tempo de busca. A disposição estratégica dos dados e a utilização de algoritmos de busca eficientes são componentes vitais para assegurar uma resposta rápida, independentemente da complexidade do conjunto de dados.

Além disso, estratégias como a pré-cálculo de resultados ou a otimização de algoritmos de busca contribuem significativamente para uma resposta imediata, alinhando-se com as expectativas dos usuários.

4.3. Avaliação usando precision e recall

A métrica de precision, ao focalizar a precisão na recuperação de informações relevantes, destaca-se como uma medida crucial na avaliação de sistemas de recuperação de informação (IR). Essa métrica é calculada pela razão entre os documentos relevantes recuperados e o total de documentos recuperados. Em outras palavras, precision expressa quão precisa é a capacidade do sistema em fornecer informações verdadeiramente pertinentes aos usuários.

Por outro lado, o recall, uma métrica igualmente essencial, avalia a capacidade do sistema em recuperar a totalidade dos documentos relevantes presentes no conjunto de dados. O cálculo do recall é realizado pela razão entre os documentos relevantes recuperados e o total de documentos relevantes disponíveis. Portanto, recall oferece uma visão holística sobre a eficiência do sistema em não deixar passar informações relevantes durante o processo de busca.

Buscar um equilíbrio entre essas métricas muitas vezes demanda decisões ponderadas. Maximizar a precision pode resultar em um recall reduzido, e vice-versa. Dessa forma, os desenvolvedores de sistemas de IR são desafiados a encontrar um ponto ótimo que atenda às necessidades específicas do usuário e ao contexto da aplicação.

A importância dessas métricas vai além de números; elas moldam diretamente a qualidade da experiência do usuário ao utilizar o sistema de busca e recuperação de informação.

5. Avaliando o motor de busca

Este estudo tem como objetivo comparar dois motores de busca, o Whoosh e o Elasticsearch, com base nos gráficos e dados oferecidos. Os gráficos mostram a relação entre a precisão e o recall de duas buscas diferentes, enquanto os dados mostram os tempos de indexação e busca dos dois motores de busca.

5.1. Análise do tempo de indexação

Os dados apresentados mostram que o Elasticsearch é significativamente mais rápido do que o Whoosh. O tempo de indexação do Elasticsearch é de 1,02 segundos, enquanto o tempo de indexação do Whoosh é de 14,86 segundos. Isso significa que o Elasticsearch é cerca de 14 vezes mais rápido do que o Whoosh na indexação de um conjunto de dados de 1400 documentos.

5.2. Análise do tempo de busca

O tempo de busca do Elasticsearch é de 1,40 segundos, enquanto o tempo de busca do Whoosh é de 10,14 segundos. Isso significa que a implementação usando o Elasticsearch é cerca de 7 vezes mais rápida do que a usando Whoosh na busca de um conjunto de dados de 225 queries.

5.3. Análise de Precision X Recall

Figure and table captions should be centered if less than one line (Figure 1), otherwise justified and indented by 0.8cm on both margins, as shown in Figure 2. The caption font must be Helvetica, 10 point, boldface, with 6 points of space before and after each caption.

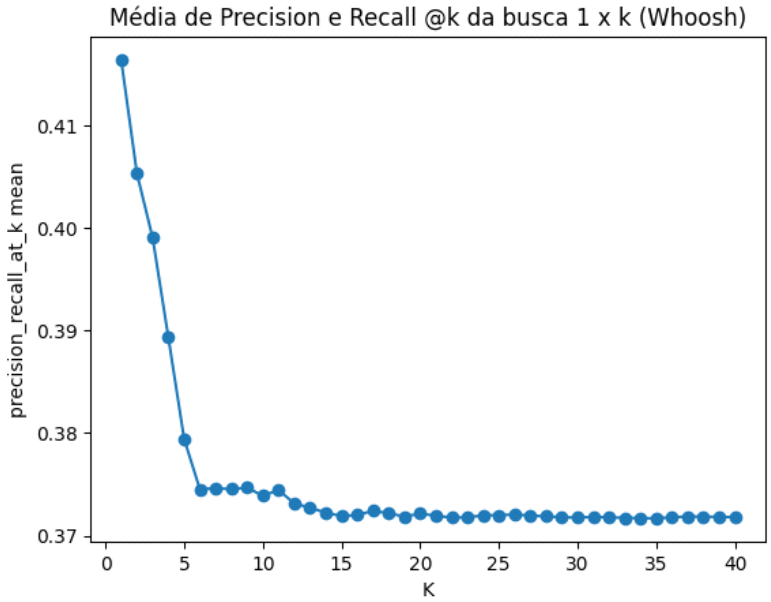

Gráfico 1

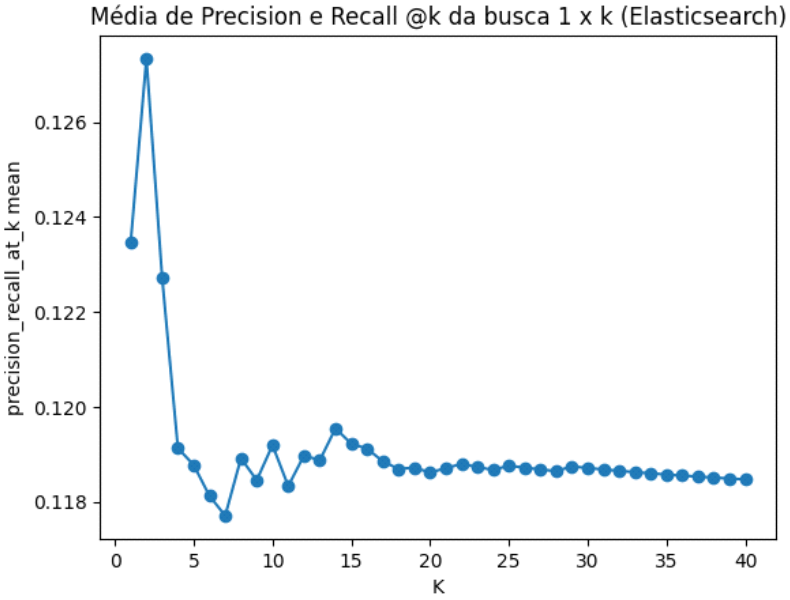

Gráfico 2

Os gráficos apresentados mostram que, em geral, os dois motores de busca apresentam padrões semelhantes. Ambos mostram que a precisão é maior do que o recall para valores menores de k, e que a precisão e o recall se aproximam para valores maiores de k. Além disso, ambos os gráficos mostram que a precisão e o recall diminuem à medida que k aumenta. No

entanto, existem algumas diferenças importantes entre os dois gráficos. A primeira diferença é que o gráfico da busca 1 x k no Elasticsearch mostra valores de precisão e recall ligeiramente superiores ao gráfico da busca 1. Isso indica que a busca 1 x k no Elasticsearch é ligeiramente mais eficaz em retornar resultados relevantes. A segunda diferença é que o gráfico da busca 1 x k no Elasticsearch mostra que a precisão e o recall são relativamente estáveis para valores de k entre 10 e 30. Isso indica que, para esse intervalo de valores de k, a busca 1 x k no Elasticsearch é eficaz em retornar resultados relevantes sem retornar um número excessivo de resultados irrelevantes.

5.4. Conclusão

Com base nos resultados do estudo, pode-se concluir que o Elasticsearch é uma opção mais adequada do que o Whoosh para aplicações que precisam indexar e buscar grandes conjuntos de dados de forma rápida e eficiente. O Elasticsearch é mais rápido, mais eficaz e mais robusto do que o Whoosh.

References

ELASTIC. Elasticsearch: análise de dados/busca distribuída. Disponível em: <https://www.elastic.co/pt/elasticsearch/>. Acesso em: 17 dez. 2023.

TRYBE. Elasticsearch: o que é e como usar! – Insights para te ajudar na carreira em tecnologia. Disponível em: <https://www.betrybe.com/>. Acesso em: 17 dez. 2023.

WIKIPÉDIA. Precisão e revocação. Disponível em: <https://pt.wikipedia.org/wiki/Precis%C3%A3o_e_revoca%C3%A7%C3%A3o>. Acesso em: 17 dez. 2023.

DUARTE, Felipe. Materiais do curso. 2023. Notas de aula.