Meta 1 - Relatório Técnico

Licenciatura em Engenharia Informática
Sistemas Distribuídos

Carlos Soares 2020230124, uc2020230124@student.uc.pt
Miguel Machado 2020230124, uc2020230124@student.uc.pt

March 23, 2025

1. Introdução

Este relatório descreve o desenvolvimento de um sistema distribuído de indexação e pesquisa de páginas web, utilizando Java RMI. O projeto visa aplicar os conceitos de tolerância a falhas, replicação, concorrência e modularidade num ambiente de redes e sistemas distribuídos.

2. Objetivos do Projeto

- Permitir ao utilizador pesquisar termos e obter páginas ordenadas por relevância;
- Indexar conteúdos web recursivamente, a partir de links inseridos;
- Consultar backlinks e estatísticas (termos mais pesquisados, tempo médio);
- Assegurar tolerância a falhas e balanceamento entre múltiplos servidores (barrels);
- Permitir execução paralela de crawlers;
- Armazenar dados persistentemente mesmo após falhas ou reinícios.

3. Componentes do Sistema

O sistema é composto pelos seguintes módulos:

- **Cliente (SearchClient)**: Interface textual onde o utilizador pode:
  - Pesquisar termos;
  - Consultar estatísticas e backlinks;
  - Adicionar links à fila central de indexação.
- **Gateway (SearchGateway)**: Encaminha pedidos do cliente para barrels ativos, com tolerância a falhas e balanceamento.
- **Barrels (IndexStorageBarrel)**: Servidores que armazenam os índices invertidos e processam pesquisas. Operam com dados replicados e oferecem persistência via ficheiros.
- **Fila Central (CentralURLQueue)**: Interface RMI responsável por armazenar e fornecer links aos crawlers.
- **Crawler (WebCrawler)**: Consome links da fila, faz scraping da página com JSoup, extrai texto e links e envia para um barrel.
- **LinkAdder**: Aplicação de linha de comando usada para adicionar links à fila.

4. Funcionamento Geral

4.1. Fluxo de Indexação

1. Utilizador insere um link pelo menu (opção 4);
2. O link é adicionado à *CentralURLQueue*;
3. Um *WebCrawler* ativo consome o link da fila;
4. O crawler extrai o texto e os links da página com a biblioteca JSoup;
5. A informação é enviada para um barrel disponível;
6. O barrel atualiza os índices e armazena backlinks.

4.2. Fluxo de Pesquisa

1. Utilizador pesquisa um termo pelo cliente;
2. A pesquisa é encaminhada à *SearchGateway*;
3. A gateway escolhe um barrel disponível e envia o pedido;
4. O barrel devolve os resultados ordenados por backlinks;
5. O cliente apresenta os resultados 10 por página.

4.3. Outras Funcionalidades

- Estatísticas de uso: top 10 termos e tempo médio de resposta;
- Consulta de backlinks de uma URL;
- Interface tolerante a falhas: o sistema funciona mesmo que um barrel falhe;
- Persistência com ficheiros .ser e .txt;
- Crawlers paralelos podem ser iniciados manualmente.

5. Tolerância a Falhas e Confiabilidade

- A gateway tenta múltiplos barrels em ordem aleatória;
- Se um barrel falhar, o outro é tentado automaticamente;
- Os barrels são réplicas: ambos armazenam os mesmos dados;
- A fila central evita duplicação de indexação entre crawlers;
- Dados são recuperados ao reiniciar qualquer barrel;
- Cada componente é modular e independente.

6. Testes Realizados

- Adição de links via LinkAdder → verificação na fila e no WebCrawler;
- Verificação da indexação no terminal dos barrels;
- Pesquisa de termos e análise dos resultados;
- Consulta de backlinks funcionais;
- Simulação de falha de barrel → continuidade do sistema;
- Reinício de barrel → dados anteriores carregados corretamente;
- Múltiplos crawlers → sem duplicação de tarefas.

7. Divisão de Trabalho

- **Carlos Soares**:
- **Miguel Machado**:
- **Ambos**:

8. Conclusão

O sistema cumpre todos os requisitos funcionais e técnicos da Meta 1. Foi testado com sucesso em cenários de uso real e simulação de falhas. A comunicação via RMI, a modularidade dos componentes, e a facilidade de extensão fazem deste projeto uma base sólida para fases futuras.