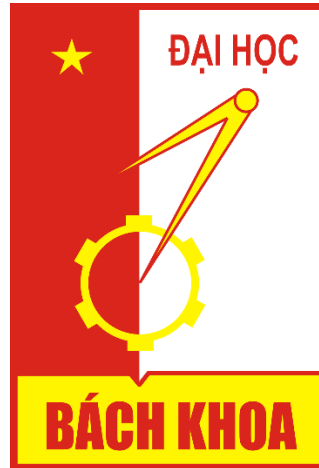


ĐẠI HỌC BÁCH KHOA HÀ NỘI
KHOA TOÁN – TIN



NHÓM 12 – BÁO CÁO TIẾN ĐỘ BÀI TẬP LỚN

HỌC PHẦN: HỆ THỐNG VÀ MẠNG MÁY TÍNH (MI4060)

CHỦ ĐỀ: Distributed Computing

ĐỀ TÀI: Thiết kế hệ thống hiệu suất cao dựa trên
trình thu thập dữ liệu thông tin web phân tán – *Distributed Web Crawler*

Giảng viên hướng dẫn: TS. Ngô Thị Hiền

Hà Nội, Tháng 11/2024

MỤC LỤC

PHẦN 1. LÝ THUYẾT CƠ BẢN

1.1. Giới thiệu về Distributed Web Crawler

- 1.1.1. Tổng quan về Distributed Web Crawler
- 1.1.2. Chức năng của Distributed Web Crawler
- 1.1.3. Ứng dụng thực tiễn của Distributed Web Crawler

1.2. Hiệu suất và hiệu suất cao

- 1.2.1. Lý thuyết hiệu suất
 - a) Định nghĩa hiệu suất
 - b) Tiêu chí đánh giá hiệu suất
- 1.2.2. Hiệu suất cao
 - a) Định nghĩa hiệu suất cao
 - b) Tiêu chuẩn và yêu cầu của hiệu suất cao
 - c) Công nghệ và công cụ hỗ trợ hiệu suất cao

1.3. Yêu cầu hiệu suất trong Distributed Web Crawler

- 1.3.1. Distributed Web Crawler và vấn đề hiệu suất
- 1.3.2. Tầm quan trọng của hiệu suất cao trong hệ thống phân tán
- 1.3.3. So sánh hiệu quả sử dụng và yêu cầu hệ thống

1.4. Chiến lược tối ưu hóa hiệu suất cho Distributed Web Crawler

- 1.4.1. Phân phối công việc thông minh (Smart Task Distribution)
- 1.4.2. Cân bằng tải (Load Balancing)
- 1.4.3. Giảm thiểu độ trễ (Latency Optimization)
- 1.4.4. Tối ưu hóa tài nguyên hệ thống (System Resource Optimization)
- 1.4.5. Kiến trúc lưu trữ dữ liệu hiệu quả
 - a) Kiến trúc lưu trữ dữ liệu hiệu quả
 - b) Phân tán dữ liệu và lưu trữ theo phân vùng (Sharding)

c) *Lưu trữ dữ liệu thu thập theo batch*

1.4.6. Xây dựng chiến lược chịu lỗi (Fault Tolerance)

1.4.7. Tối ưu hóa thuật toán thu thập dữ liệu

a) *Thuật toán thu thập dữ liệu ưu tiên (Priority-Based Crawling)*

b) *Thuật toán phòng tránh lặp lại (De-duplication)*

c) *Thuật toán tối ưu hóa truy cập nhiều trang liên quan*

1.4.8. Ứng dụng các công nghệ và công cụ hỗ trợ hiệu suất

PHẦN 2. THỰC HÀNH

2.1. Thiết kế hệ thống với trọng tâm hiệu suất cao

2.2. Triển khai và cài đặt hệ thống tối ưu hiệu suất

2.3. Thử nghiệm và đo lường hiệu suất

2.4. Phân tích kết quả và tối ưu hóa

≈ HẾT ≈