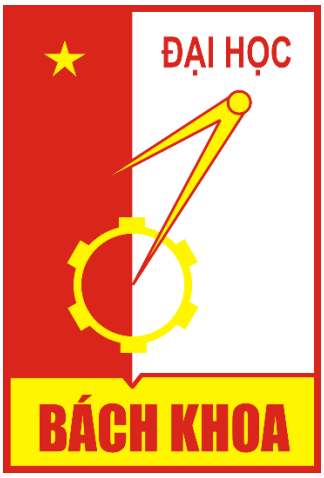
ĐẠI HỌC BÁCH KHOA HÀ NỘI

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

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