

Proyecto Final

Curso de Sistemas Operativos y Laboratorio

Análisis Paralelo de Sentimiento en Reseñas de Productos

Juan Esteban Ortiz

Israel Velez

Resumen

Este proyecto desarrolla una aplicación de análisis de sentimientos utilizando programación paralela en C, aplicando la librería `pthread.h` para implementar hilos de procesamiento simultáneo. Se analizarán reseñas de productos para identificar la frecuencia de palabras positivas y negativas. La implementación permite procesar múltiples reseñas en paralelo, mejorando el rendimiento y reduciendo el tiempo de análisis en comparación con una versión secuencial. Este enfoque tiene aplicaciones prácticas en la minería de datos y el procesamiento de lenguaje natural, relevantes en el contexto de comercio electrónico y redes sociales.

Introducción

- **¿Cuál es la necesidad y/o problema que aborda el desafío seleccionado?**

El análisis de sentimientos permite entender mejor las opiniones de los clientes en sectores como el comercio electrónico, donde una gran cantidad de reseñas de usuarios requiere análisis rápido y eficiente. Este proyecto busca abordar la necesidad de una solución que pueda analizar el sentimiento de grandes volúmenes de datos en tiempo

real o casi en tiempo real, mejorando así la capacidad de las empresas para responder a las opiniones de sus clientes.

- **¿Por qué es importante el desarrollo de este desafío en el contexto tecnológico actual?**

En el entorno actual, la velocidad y la precisión del análisis de datos son cruciales para que las empresas respondan rápidamente a las tendencias y preferencias de los usuarios. La implementación de procesamiento paralelo en el análisis de sentimientos reduce significativamente el tiempo de ejecución, lo cual es fundamental cuando se manejan grandes volúmenes de datos en tiempo real, como en el caso de las plataformas de comercio electrónico y redes sociales.

Antecedentes o marco teórico

¿Cuáles son los principales aspectos teóricos necesarios para comprender el desafío y llevarlo a cabo?

- **Programación Paralela y Concurrencia:** La programación paralela distribuye tareas entre múltiples hilos de ejecución para reducir el tiempo de procesamiento. En este caso, `pthread.h` permite la creación y manejo de hilos que procesan reseñas de forma concurrente.
- **Análisis de Sentimientos y Procesamiento de Lenguaje Natural (NLP):** Este proyecto emplea técnicas básicas de NLP para analizar palabras específicas en un texto y categorizar las reseñas en base a sentimientos positivos o negativos.
- **Condiciones de Carrera y Sección Crítica:** La programación concurrente puede generar conflictos si varios hilos acceden simultáneamente a datos compartidos. Las secciones críticas y los mutexes ayudan a evitar condiciones de carrera, garantizando que el acceso a los recursos compartidos sea seguro.

¿Qué relación tiene esta teoría con los temas del curso de Sistemas Operativos?

Este proyecto aplica conceptos fundamentales de sistemas operativos, incluyendo la gestión de hilos y la concurrencia, que se estudian en teoría y en laboratorio. `pthread.h` permite una implementación práctica de la creación y sincronización de hilos, y el manejo de secciones críticas mediante mutexes se relaciona directamente con la teoría de exclusión mutua y condiciones de carrera en los sistemas operativos.

Objetivos (principal y específicos)

Objetivo Principal

Implementar un sistema paralelo para el análisis de sentimientos en reseñas de productos, usando programación con hilos para mejorar la eficiencia en comparación con una implementación secuencial.

Objetivos Específicos

- Diseñar un programa en C que cargue múltiples reseñas de un archivo y las procese en paralelo usando `pthread.h`.
- Implementar el conteo de palabras positivas y negativas en cada reseña, categorizando su sentimiento.
- Evaluar el rendimiento comparativo entre la versión paralela y la secuencial del análisis de sentimientos.

Metodología

Herramientas Principales

- **Lenguaje C y `pthread.h`:** para la programación concurrente y el manejo de hilos.
- **Editor de Texto o Entorno de Desarrollo** (Visual Studio Code) para la implementación y compilación del código en C.
- **Cronómetro o función `clock()` de C** para medir el rendimiento de las versiones secuencial y paralela.

Actividades Necesarias

1. **Investigación y diseño:** Investigar cómo `pthread.h` permite la creación y manejo de hilos en C y diseñar el flujo general del programa.
2. **Preparación de los datos:** Crear archivos de texto con reseñas y listas de palabras positivas y negativas.
3. **Desarrollo del programa:** Implementar funciones para cargar reseñas, cargar palabras, y realizar el análisis paralelo.
4. **Pruebas unitarias:** Probar la funcionalidad del programa con un subconjunto de datos para asegurar que funciona correctamente.
5. **Medición del rendimiento:** Ejecutar la versión secuencial y paralela del programa y registrar los tiempos de ejecución para análisis de rendimiento.

6. **Análisis y documentación:** Analizar los resultados y documentar los hallazgos en el informe final.

-

Diseño de Experimentos

Validación de la funcionalidad y rendimiento

Para validar el desempeño del sistema, se medirán los tiempos de ejecución de las versiones secuencial y paralela con distintos tamaños de datos. Además, se verificará la precisión en el conteo de palabras de sentimiento para asegurar que ambas versiones arrojen resultados consistentes. Se planea medir:

- **Tiempo de ejecución** en distintos tamaños de datos (p.ej., 100, 1000, 5000 reseñas).
- **Precisión en el conteo:** verificar que ambas versiones contabilicen correctamente las palabras positivas y negativas.
- **Eficiencia en escalabilidad:** evaluar el rendimiento al aumentar el número de hilos.