UNSL

Universidad Nacional de San Luis
Facultad de Ciencias Físico, Matemáticas y Naturales
Departamento de Informática

Tesis
Maestría en Ingeniería de Software

Proyecto e Implementación de un Algoritmo Distribuido para Búsquedas por Similitud Basado en la Estructura D-Index

Lic. Norma Beatriz Perez

Asesores Científicos:
Director: Fernando Magno Quintão Pereira, PhD.
Co-Director: Dir. Mario Marcelo Berón

San Luis - Argentina

12 Noviembre, 2013

Agradecimientos

Ante todo he de agradecer a mi familia, Sofi, Sylvain y Adeline por entender mi ausencia, acompañandome a transitar esta etapa con su cariño y paciencia.

Agradezco la colaboración de mi director de tesis Fernando Magno Quintão Pereira y co-director Mario Marcelo Berón por permitirme formar parte de sus grupos de tesistas, por el profesionalismo, orientación, por todo lo que he aprendido de ellos y todas las oportunidades de crecimiento que me han brindado. Gracias por siempre haber encontrando las palabras precisas para invitarme a seguir.

Agradezco al Laboratorio *Speed* del Departamento de Ciências da Computação (DCC) de la Universidade Federal de Minas Gerais (UFMG), Belo Horizonte - Brasil, por permitirme usar sus servidores. A Bruno Rocha Coutinho, Ana Paula de Carvalho, Thatyene Louise Alves de Souza Ramos y Rodrigo Silva Oliveira por asistirme e instruirme tanto en el manejo de los clusters como la infraestructura Watershed, sin su apoyo me hubiera resultado imposible terminar mi trabajo.

También deseo expresar mi gratitud a: aquellos profesores, compañeros de trabajo y amigos, cuya personalidad, métodos y amistad han influido en mí, ayudándome a crecer. Especialmente a los profesores Carlos Camarão de Figueiredo, Antonio Alfredo F. Loureiro, Clarindo Isaías Pereira da Silva e Pádua, José Monteiro da Mata y José Marcos Silva Nogueira quienes me acompañaron en mi etapa en la UFMG; a Pablo, Paula, Paola, Alejandro, Dario, Javier, Vanesa, Jesús, Nilda, Gabriela, Jorge, Junio, Natiela, Natalia, Roxana, Maria José, Augusto con quienes he compartido horas y horas de estudio, de quiénes he recibido siempre palabras de aliento y confianza y al grupo de Compresión de Programa de la Universidad Nacional de San Luis (Enrique Miranda, Hernán Bernardis, Matias Truglio, Edgardo Bernadis, Mariano Luza y su director Mario Berón), por la ayuda brindada en una etapa de mi carrera.

Índice general

Índice de Figuras viii

Índice de Tablas xi

Índice de Algoritmos xii

1 Conceptos Preliminares 1

1.1 Introducción . . . 2
1.2 Estructuras de Datos . . . 3
1.3 Modelos e Infraestructuras de Computación Paralela . . . 6
1.4 Presentación del Problema . . . 9
1.5 Ejemplo de Aplicación . . . 11
1.6 La Hipótesis . . . 12
1.7 Solución Propuesta del Problema . . . 12
1.8 Estructura de la Disertación . . . 14

2 Modelos e Infraestructuras de Computación Paralela 17

2.1 Modelos de Computación Paralela . . . 18
2.1.1 Parallel Random Access Machine . . . 19
2.1.2 Filter-Stream . . . 20
2.1.3 Bulk-Synchronous Parallel . . . 22
2.1.4 Modelo LogP . . . 24
2.1.5 MapReduce . . . 25
2.2 Infraestructura de Programación Paralela Basada en Filter-Stream . 26
2.2.1 DataCutter . . . 26
2.2.2 Anthill . . . 26
2.2.3 Watershed . . . 27

2.3 Notas y Comentarios 29

3 Sustento Teórico y Estado del Arte de Búsquedas por Similitud 31

3.1 Modelo de los Espacios Métricos 32
3.2 Funciones de Distancias 32
3.3 Tipos de Búsquedas en los Espacios Métricos 34
3.4 Dimensionalidad 36
3.5 Indexación de Algoritmos 38
3.6 Clasificación de los Algoritmos 39
3.7 Principios de Particionamiento 41
3.7.1 Particionamiento por Bolas 41
3.7.2 Hiper-plano Generalizado 43
3.7.3 Particionamiento de Exclusión de la Parte Media 43
3.8 Notas y Comentarios 44

4 Infraestructura Watershed 47

4.1 Visión General 47
4.2 Semántica de Watershed 48
4.3 El Ambiente de Watershed 51
4.4 Arquitectura 54
4.4.1 Etapa de Programación 56
4.4.2 Etapa de Control 60
4.4.3 Etapa de Comunicación 62
4.5 Un Ejemplo de Uso de Watershed 63
4.6 Notas y Comentarios 71

5 D-Index 73

5.1 La estructura D-Index 74
5.2 D-Index en un Ejemplo 74
5.2.1 Clustering a través de Particionamiento Separables 78
5.3 Filtrado por Pivotes Aplicado al D-Index 84
5.3.1 Arquitectura del D-Index 87
5.3.2 Algoritmo de Inserción 88
5.3.3 Algoritmo de Búsqueda por Rango 89

6 Técnicas de Paralelización Propuestas en este Trabajo 93
6.1 Paralelización del D-Index 94
6.2 Construcción del Índice 95
6.3 Búsquedas sobre el Índice 98
6.4 Construcción y Búsqueda del D-Index en Watershed 102
6.5 Notas y Comentarios 104

7 Experimentos y Resultados 107
7.1 Fundamentos de los Experimentos 107
7.2 Ambiente Experimental 108
7.3 Resultados Experimentales 110
7.4 Técnica Naïve 111
7.4.1 Tiempo de Respuesta - Histograma de COLOR 111
7.4.2 Tiempo de Respuesta - Imágenes de la NASA 112
7.4.3 Speed-up - Histograma de COLOR 113
7.4.4 Speed-up - Imágenes de la NASA 113
7.4.5 Throughput - Histograma de COLOR 114
7.4.6 Throughput - Imágenes de la NASA 115
7.5 Técnica Local 116
7.5.1 Tiempo de Respuesta - Histograma de COLOR 116
7.5.2 Tiempo de Respuesta - Imágenes de la NASA 117
7.5.3 Speed-up - Histograma de COLOR 117
7.5.4 Speed-up - Imágenes de la NASA 118
7.5.5 Throughput - Histograma de COLOR 119
7.5.6 Throughput - Imágenes de la NASA 120
7.6 Técnica Global 120
7.6.1 Tiempo de Respuesta - Histograma de COLOR 121
7.6.2 Tiempo de Respuesta - Imágenes de la NASA 121
7.6.3 Speed-up - Histograma de COLOR 122
7.6.4 Speed-up - Imágenes de la NASA 122
7.6.5 Throughput - Histograma de COLOR 123
7.6.6 Throughput - Imágenes de la NASA 124
7.7 Conclusión de los Experimentos 129

8 Conclusiones y Trabajos Futuros 131

A Notación utilizada **137**

B Lista de Acrónimos **139**

Bibliografía **141**

Índice de figuras

2.1 Arquitectura de un sistema procesamiento *Data-Stream*. 19
2.2 Visión del programador del modelo *filter-stream*. 21
2.3 Visión de la infraestructura del modelo *filter-stream*. 21
2.4 Superpaso de BSP. 22
2.5 Modelo arquitectónico de la máquina BSP. 23
2.6 Composición de procesamiento en el modelo *filtro-stream*. 28

3.1 Búsqueda por rango: $(q, r)_d = \{x_2, x_7, x_9, x_{11}\}$. 35
3.2 Búsqueda de los 4-NN $=\{x_2, x_7, x_9, x_{11}\}$. 36
3.3 Un histograma de distancias para un espacio métrico de dimensión baja (izquierda) y de dimensión alta (derecha). 37
3.4 Modelo general utilizado para indexación y consulta en espacios métricos. 38
3.5 Taxonomía de los algoritmos. 41
3.6 Particionamiento por Bolas. 43
3.7 Hiper-plano Generalizado. 44
3.8 Particionamiento Exclusión de la parte media. 45

4.1 La semántica de Watershed. 49
4.2 Función auxiliar utilizada para definir la semática de Watershed. . . 50
4.3 Arquitectura del ambiente de procesamiento Watershed. 55
4.4 Diagrama de operación de la consola de la infraestructura Watershed. . . 61
4.5 Procesamiento interno de los filtros, en la aplicación del ejemplo. . . 64
4.6 Topología de los filtros de flujo del ejemplo presentado. 70

5.1 Ejemplo del D-Index 75
5.2 D-Index: Jerarquía por buckets. 77

5.3 Paticionamiento. 80
5.4 Combinación de dos bps. 81
5.5 Ejemplo de cómo se utiliza la función G. 83
5.6 Cómputo de la función de distancia. 85
5.7 Cómputo de la función de distancia. 85
5.8 Ejemplo: Inserción de objectos en el D-Index. 88
5.9 Ejemplo: búsqueda por rango 90

6.1 Técnica de construcción naïve. 96
6.2 Técnica de construcción local. 97
6.3 Técnica de construcción local. 98
6.4 Proceso de búsqueda para el enfoque de paralelización Naïve. 99
6.5 Proceso de búsqueda para el enfoque de paralelización Local. 100
6.6 Proceso de búsqueda para el enfoque de paralelización Global. 101
6.7 Filtros que componen el proceso de construcción y búsqueda del D-Index en Watershed 104

7.1 Tiempo por consulta, espacio de Histogramas de COLOR. 111
7.2 Tiempo por consulta, espacio de Imágenes de la NASA. 112
7.3 Speed-up, espacio de Histogramas de COLOR. 113
7.4 Speed-up, espacio de Imágenes de la NASA. 114
7.5 Throughput, espacio de Histogramas de COLOR. 115
7.6 Throughput, espacio de Imágenes de la NASA. 115
7.7 Tiempo por consulta, espacio de Histogramas de COLOR. 116
7.8 Tiempo por consulta, espacio de Imágenes de la NASA. 117
7.9 Speed-up, espacio de Histogramas de COLOR. 118
7.10 Speed-up, espacio de Imágenes de la NASA. 119
7.11 Throughput, espacio de Histogramas de COLOR. 119
7.12 Throughput, espacio de Imágenes de la NASA. 120
7.13 Tiempo por consulta, espacio de Histogramas de COLOR. 121
7.14 Tiempo por consulta, espacio de Imagénes de la NASA. 121
7.15 Speed-up, espacio de Histogramas de COLOR. 122
7.16 Speed-up, espacio de Imágenes de la NASA. 123
7.17 Throughput, espacio de Histogramas de COLOR. 123
7.18 Throughput, espacio de Imágenes de la NASA. 124

Índice de cuadros

3.1 Resumen de las complejidades promedio de las estructuras existentes más conocidas. 42

7.1 Radios para la colección Histograma de COLOR. 110
7.2 Radios para la colección Imágenes de la NASA. 111
7.3 Colección Histograma de COLOR: valores del *speed-up* obtenidos experimentalmente según la técnica aplicada para diferentes porcentajes de recuperación. 126
7.4 Colección Imágenes de la NASA: valores del *speed-up* obtenidos experimentalmente según la técnica aplicada para diferentes porcentajes de recuperación. 127
7.5 Colección Histograma de COLOR: valores del hroughput obtenidos experimentalmente según la técnica aplicada para diferentes porcentajes de recuperación. 128
7.6 Colección Imágenes de la NASA: valores del *hroughput* obtenidos experimentalmente según la técnica aplicada para diferentes porcentajes de recuperación. 129
7.7 Eficiencia de cada colección respecto a las diferentes técnicas para $P = 10$ procesadores. 130

Lista de algoritmos

4.1 **Archivo DTD de Watershed** 52
4.2 **Achivo de configuración de Watershed** 53
4.3 **Archivo XML de configuración** 54
4.4 **Archivo XML de configuración del filtro search_local** 57
4.5 **Filtro Reader (reader.cc)** 65
4.6 **Filtro de configuración XML (reader.xml)** 66
4.7 **Filtro Adder (adder.cc)** 67
4.8 **Archivo de configuración XML Adder (adder.xml)** 68
4.9 **Filtro Writer (writer.cc)** 69
4.10 **Archivo de Configuración XML Writer(writer.xml).** 70
5.1 Algoritmo de búsqueda con distancia por pivotes **D** 87
5.2 Algoritmo de inserción del D-Index 89
5.3 Algoritmo búsqueda por rango 91