



UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

Estudio de tamaños Mínimos de Rings y Order Graphs para Joins Óptimos

PROPUESTA DE TEMA DE MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL EN COMPUTACIÓN

Javier Oliva

MODALIDAD:
Memoria

PROFESOR GUÍA:
Gonzalo Navarro

SANTIAGO DE CHILE
2025

1. Introducción

En el ámbito de las bases de datos, las consultas de tipo *join* son fundamentales para la extracción de información a partir de múltiples tablas. Dado que su uso es extenso sobre bases de datos masivas, el desarrollo de algoritmos eficientes en tiempo de ejecución sobre representaciones eficientes en espacio se vuelve relevante. En este contexto, el estudio de algoritmos eficientes en tiempo de ejecución en el peor caso, o *Worst-case optimal* (wco) [2], adquiere un rol especialmente importante pues entrega garantías para cualquier consulta que se realice.

Varios estudios buscan mejorar el tiempo de ejecución de algoritmos wco, pero pocos buscan mejorar el espacio requerido por ellos. En particular, los algoritmos wco actuales necesitan un espacio exponencial con respecto a la aridad de la relación, por lo que se vuelven poco prácticos rápidamente para aridades pequeñas. Por ello desarrollar algoritmos wco que utilicen menos espacio en trade off con respecto a la velocidad de ejecución puede significar el posible de algoritmos wco para aridades más grandes.

De aquí viene el Ring [1], una estructura de datos que permite algoritmos wco con un uso muy reducido de espacio. Un Ring se puede representar como un orden cíclico bidireccional de columnas, por ende tiene tamaño igual a la aridad de la relación, donde el caso de aridad 3 necesita solo un ring para implementar un algoritmo wco mientras que el caso de aridad 4 necesita dos rings para cubrir todos los casos. El número de rings necesarios para implementar algoritmos wco es desconocido en el caso general y conocido para aridades menores o iguales a 6 [1]. En la actualidad solo se conocen cotas inferiores y superiores no ajustadas para el mínimo número de rings necesarios para aridades mayores a 6.

Además de los Rings están los Order Graphs [1], una estructura que generaliza los Rings en el sentido de que un conjunto de Rings se puede representar como un Order Graph, pero un Order Graph no necesariamente se puede representar como un conjunto de Rings. Los Order Graph completos también permiten algoritmos wco y, al contener los Rings, evidentemente en el peor caso serán tan buenos como los Rings en espacio. Por ejemplo, se sabe que para aridad igual a 5 hay un Order Graph completo que utiliza menos espacio que el conjunto de Rings óptimo, mientras que para aridad 3, 4 y 6 un Order Graph completo óptimo en espacio utiliza el mismo espacio que un conjunto óptimo de Rings. Para aridades mayores se desconoce si los Order Graph completos óptimos son estrictamente mejores que los conjuntos de Rings óptimos en espacio o no.

Los Rings y Order Graphs son un área de estudio reciente y, por lo tanto, poco desarrollada. Como se ve, se encuentra un estudio más completo para casos de aridad ≤ 6 mientras que para aridades mayores solo se tienen cotas poco ajustadas sobre los tamaños de los óptimos. El caso de aridad 3 muestra resultados[1] que indican la utilidad práctica de estas estructuras, por lo que se espera que para aridades mayores también sea el caso. El trabajo presente busca el desarrollo de esta área de estudio para aridades ≥ 7 que va desde encontrar mejores Order Graphs completos o conjuntos de Rings que los conocidos, a encontrar cotas mejores que las conocidas en el caso general de aridad d .

2. Situación Actual

Un Order Graph de dimensión d es un grafo dirigido etiquetado $\mathcal{G}(\mathcal{V}, \mathcal{E})$ donde $\mathcal{V}$ es un subconjunto de los $d!$ posibles órdenes de $\{1, 2, \dots, d\}$ y un arco de un nodo Π a Π' , etiquetado $j \in \{1, 2, \dots, d\}$, existe si Π' se obtiene de Π moviendo el elemento j al frente.

Por ejemplo con $d = 4$ y nodos $\Pi = 2143$ y $\Pi' = 1243$ entonces el arco $2143 \xrightarrow{1} 1243$ puede existir en un order graph, tiene etiqueta 1 pues mueve el 1 al frente (inicio) del orden.

Un Order Graph de dimensión d es *completo* si, para cada subconjunto $S \subsetneq \{1, 2, \dots, d\}$ y todo $x \in \{1, 2, \dots, d\} \setminus S$, existe un camino dirigido de largo $|S| + 1$ en $\mathcal{G}$ cuyas etiquetas contienen exactamente los elementos de $S \cup \{x\}$ en algún orden, tal que la etiqueta x se encuentra al inicio o final del camino.

Un Ring de dimensión d es un ciclo de largo d con las etiquetas $\{1, 2, \dots, d\}$ en algún orden. Noten que un Ring es un Order Graph.

Decimos que un Order Graph completo tiene tamaño óptimo si no hay Order Graph completo que tenga menos arcos.

Para dar un ejemplo, se muestran los 8 Order Graph óptimos [1] de dimensión 4.

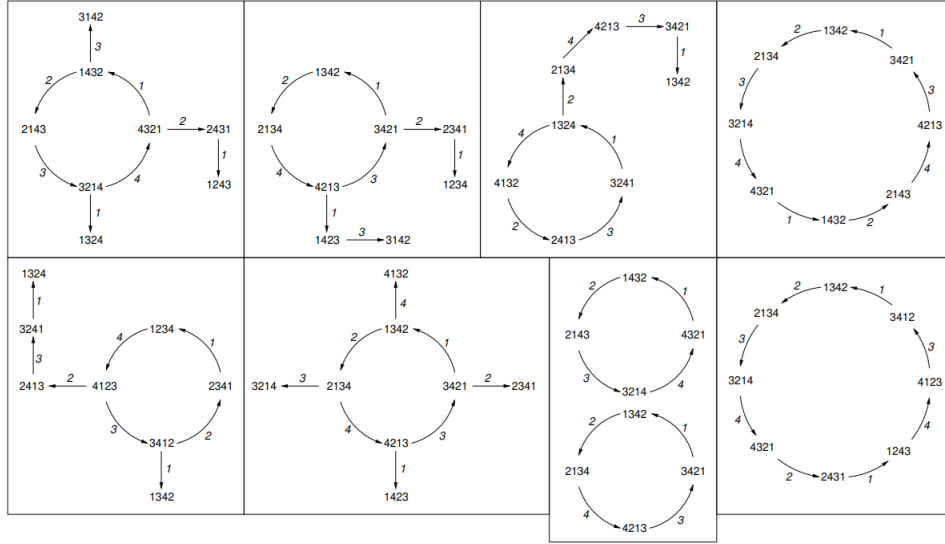


Figura 1: Order Graphs completos de dimensión 4.

Donde el tercer grafo de la segunda fila corresponde a un conjunto de Rings que permite algoritmos wco, mostrando que en este caso de dimensión 4 los Order Graphs no son estrictamente mejores en espacio.

Los resultados actuales se presentan en la siguiente tabla, considerando el número de nodos como unidad. En este caso solo se estudian los Order Graph completos que consisten de un solo gran ciclo, como los que se muestran en la última columna de la figura.

d/edt	Conjuntos de Rings	Order Graph completos	Cota inferior	Cota superior
2	2	2	2	2
3	3	3	3	3
4	8	8	8	8
5	25	20	20	20
6	42	42	42	42
7	84	84	70	84
8	200	209	168	200

La cota superior claramente corresponde al mínimo tamaño encontrado que corresponda a un conjunto de Rings que permite algoritmos wco u Order Graph completos que consistan de un solo gran ciclo, mientras que la cota inferior se obtiene en base a argumentos combinatoriales. Los valores para Conjuntos de Rings y Order Graph completos corresponden al menor tamaño del cual se conoce una estructura de este tipo, no necesariamente óptima.

Se puede ver que para dimensiones 7 y 8 no hay cotas ajustadas y no se sabe si lo encontrado hasta el momento es óptimo o no, mientras que para dimensiones más grandes el rango de búsqueda entre cota inferior y superior es muy grande. Que estas cotas no sean ajustadas implica que puede haber mejoras en espacio para soportar algoritmos wco.

Todos los resultados teóricos pueden encontrarse en el paper [1] que contiene los resultados del estado del arte actual. Estos resultados se obtuvieron mas que nada para mostrar ejemplos y posibles extensiones del paper utilizando una computadora normal con algoritmos secuenciales. Por lo tanto, obtener resultados mejores es altamente probable.

3. Objetivos

Objetivo General

El objetivo del presente trabajo es mejorar las cotas conocidas para tamaños óptimos de Order Graphs completos y conjuntos de Rings que permitan algoritmos wco para distintas cantidades de atributos con el fin de determinar si existen soluciones mejores a las actuales o no.

Objetivos Específicos

En lo que sigue se entiende como tiempo razonable un intervalo de tiempo menor a una semana.

1. Diseñar algoritmos de búsqueda exhaustiva y no exhaustiva para mejorar cotas existentes para Rings.
2. Diseñar algoritmos de búsqueda exhaustiva y no exhaustiva para mejorar cotas existentes para Order Graphs.

3. Implementar los algoritmos anteriores de forma eficiente para que logren ejecutarse en un tiempo razonable.
4. Obtener resultados que mejoren las cotas actuales para estas estructuras.

Evaluación

Se evaluará en base a los algoritmos y las técnicas utilizadas para encontrar óptimos. El desarrollo teórico del algoritmo, que incluye el análisis de complejidad temporal y espacial, además de las demostraciones de su funcionamiento o de que tienden a un óptimo con respecto al tiempo de ejecución. Aunque no se logre la obtención de cotas, el desarrollo teórico de estos algoritmos sumarán a futuras investigaciones.

Otro punto de evaluación es el desarrollo de resultados sobre las cotas actuales. Si se lograron encontrar mejores cotas / cotas ajustadas para dimensiones mayores o iguales a 7. Si se logra mejorar alguna cota se determina como logrado el objetivo general, si no se determina como parcialmente logrado en función del desarrollo de algoritmos.

4. Solución Propuesta

La etapa principal consiste en el desarrollo de un algoritmo de generación de estas estructuras para familiarizarse y estudiar sobre ellas, pues el uso de combinatoria puede determinar que existen estructuras óptimas que cumplan ciertas restricciones, permitiendo no buscar sobre el espacio completo. Una vez obtenido, observar patrones en las estructuras y conjeturar sobre la forma de las estructuras óptimas o sobre optimizaciones en el momento de generarlas. Se utilizarán técnicas de análisis de algoritmos avanzados junto con combinatoria y optimización combinatorial.

Hay varias técnicas distintas que se pueden probar para el desarrollo del trabajo. Lo primero es que no se está restringido a algoritmos secuenciales, se pueden utilizar algoritmos paralelos que puedan implementarse en tarjetas gráficas lo que aceleraría el tiempo de ejecución por grandes constantes. Por lo que el estudio y desarrollo de estos tipos de algoritmos es parte de lo propuesto y se espera obtener 1 a implementar.

Luego, la investigación sobre algoritmos de optimización combinatorial pueden permitir obtener reducciones de ciertas partes del problema en las cuales se puedan utilizar estos algoritmos ya estudiados previamente. Estos algoritmos no se restringen a algoritmos que encuentran un óptimo, podemos utilizar algoritmos que con respecto al tiempo tiendan a obtener resultados cada vez más cercanos al óptimo. En una primera instancia se estudiará sobre la factibilidad de Descenso de Gradiente y Simulated Annealing.

Además, se investigará sobre el uso de algoritmos secuenciales que logren una búsqueda completa para obtener el óptimo, sin importar el tiempo que tomen, si toman un tiempo mayor a una semana entonces se corta la ejecución.

5. Plan de Trabajo

1. Estudio general de las estructuras Ring y Order Graph. (2 semanas).

En esta etapa se desarrollarán algoritmos de búsqueda exhaustiva sin importar la complejidad temporal o espacial de la implementación con el objetivo de observar distintas soluciones del problema y desarrollar intuición. Esta parte es importante también para tener un punto de comparación con los siguientes algoritmos, para verificar correctitud en casos pequeños.

2. Desarrollo teórico de algoritmos de búsqueda exhaustiva y no exhaustiva. Algoritmos de aproximación, aleatorios, correctos y paralelos en GPU. (2 semanas).

En esta etapa se desarrollarán de forma teórica distintos algoritmos que mejoren en algún aspecto los algoritmos de fuerza bruta donde no necesariamente se busque un óptimo, sino que una mejor solución a las presentes en el estado del arte.

3. Escritura sobre el desarrollo realizado sobre algoritmos. (1 semana). Primer hito y evaluación sobre lo avanzado.

4. Implementación de los algoritmos obtenidos en el punto anterior. (2 semanas).

En esta etapa se implementarán los algoritmos diseñados anteriormente. Esto implica correctitud y optimizaciones necesarias para obtener un tiempo de ejecución acorde a lo teórico.

5. Realizar experimentos sobre el tiempo de ejecución de los algoritmos mencionados anteriormente junto a resultados auxiliares, que permitirán el análisis de utilidad y posibles mejoras futuras de los algoritmos. (2 semanas). Segundo hito y evaluación sobre lo avanzado.

En esta etapa se experimentará con los algoritmos implementados para mejorar constantes de tiempo de ejecución con el objetivo de lograr obtener mejores soluciones a las encontradas en el estado del arte. Los resultados de los experimentos entregarán información sobre la utilidad o factibilidad de algunos algoritmos por sobre otros en el contexto de generación de Rings y Order Graphs.

6. Escritura sobre los resultados obtenidos en el informe final, incluyendo la evaluación de los resultados. Volver al paso 2 mientras no se obtenga un algoritmo lo suficientemente rápido que obtenga resultados mejores que los actuales. (3 semanas).

En esta etapa se mejoran de manera dinámica los algoritmos ya implementados en base a los resultados de los experimentos.

7. Realización de experimentos para la obtención de cotas mejores a las actuales. (1 semana). Último hito y evaluación.

En esta etapa se realizan los experimentos finales dónde se obtendrán los mejores resultados del trabajo.

8. Escritura sobre los resultados obtenidos y finalización del informe. (2 semanas).

En esta etapa se obtienen los resultados y conclusiones finales.

6. Avances

1. Se lograron desarrollar e implementar algoritmos secuenciales de búsqueda exhaustiva en CPU con `c++` para obtener soluciones óptimas de Rings y algoritmos de búsqueda

con poda para obtener soluciones óptimas de Order Graphs de tamaño n que igualan los resultados del estado del arte para $n \leq 6$. La ejecución de estos ejecutables toma menos de 10 minutos lo que permite agilizar el desarrollo. Estos algoritmos entregan el conjunto de todas las soluciones óptimas, con lo que se tiene un resultado preliminar que no solo iguala el estado del arte, sino que también permite un estudio más profundo de la estructura.

2. Se intentó desarrollar un algoritmo voraz que permitiese obtener resultados no necesariamente óptimos para valores mayores de n en Rings pero, lamentablemente, las soluciones encontradas no compiten con el estado del arte. El avance es el descarte de varias ideas alternativas para el desarrollo de algoritmos lo cual permitirá optimizar el uso de tiempo para el desarrollo del siguiente semestre.
3. Se desarrolló e implementó un algoritmo paralelo de búsqueda exhaustiva en GPU con cuda para el estudio de Rings. Si bien el ejecutable es mucho más rápido que los secuenciales desarrollados anteriormente, tampoco logra romper la barrera de $n = 7$ y se determinó que le tomaría más de un año resolver este caso. Esto muestra que el desarrollo combinatorial de estas estructuras es necesario para avanzar el estado del arte pues es altamente probable que una búsqueda exhaustiva nunca logre el objetivo.

Los algoritmos de búsqueda exhaustiva mostraron ser insuficientes para lograr el objetivo por lo que se usarán como una herramienta para el desarrollo de teoría con respecto al problema de optimización combinatorial con el que nos vemos enfrentados. Dado esto, el próximo semestre se podrá empezar la plan de trabajo desde el punto 2 que permitirá dedicar más tiempo al desarrollo teórico de algoritmos.

Referencias

- [1] Arroyuelo, Diego, Adrián Gómez-Brandón, Aidan Hogan, Gonzalo Navarro, Juan Reutter, Javiel Rojas-Ledesma y Adrián Soto: *The Ring: Worst-case Optimal Joins in Graph Databases using (Almost) No Extra Space*. ACM Transactions on Database Systems, 49(2):1–45, 2024.
- [2] Ngo, Hung Q, Ely Porat, Christopher Ré y Atri Rudra: *Worst-case optimal join algorithms*. Journal of the ACM (JACM), 65(3):1–40, 2018.