Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 12, No. 8, Mes Agosto, 2019
ISSN: 2306-2495 | RNPS: 2343
Pág. 18-26
http://publicaciones.uci.cu

Tipo de artículo: Artículo original
Temática: Soluciones Informáticas
Recibido: 19/06/2019 | Aceptado: 20/08/2019 | Publicado: 22/08/2019

Potencialidades de la Optimización Basada en Mallas Dinámicaspara resolver el problema de la fragmentación vertical de bases de datos

Potential of Dynamic Mesh Based Optimization to solve the problem of vertical database fragmentation

Yurisbel Vega Ortiz[1*], Yandielys Reyes Plano[2]

[1] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km. 2 ½, Reparto: Torrens Municipio: Boyeros Provincia: La Habana. yurisbelv@uci.cu
[2] Universidad de las Ciencias Informáticas. Carretera a San Antonio de los Baños, Km. 2 ½, Reparto: Torrens Municipio: Boyeros Provincia: La Habana. yandie@uci.cu

[*] Autor para correspondencia: acalvarez@uci.cu

Resumen

La fragmentación vertical es el proceso mediante el cual una relación es descompuesta en agrupaciones de datos en función de conjuntos disjuntos de sus atributos. Estos conglomerados de atributos deben ser generados de manera apropiada para su futura ubicación sobre una plataforma distribuida. La partición vertical reviste gran importancia debido a que defiende el principio de que aquellos datos que comúnmente son accedidos juntos deberían residir en el mismo fragmento. La partición vertical es un problema de optimización que tradicionalmente ha sido tratado con técnicas de análisis de clúster, métodos jerárquicos y heurísticas con dos enfoques fundamentales: aglomerativo o divisorio. En los últimos años, la aplicación de meta-heurística poblacional a este problema ha incrementado su énfasis y centrado las principales investigaciones. Por ello, el objetivo de este artículo es defender las potencialidades de la Optimización basada en Mallas Dinámicas como meta-heurística viable para resolver dicho problema y obtener resultados competitivos.

Palabras clave: Bases de datos distribuidas, fragmentación vertical, medida de afinidad, optimización basada en mallas dinámicas.

Abstract

Vertical fragmentation is the process by which a relationship is broken down into data groups based on disjoint sets of their attributes. These attribute clusters should be generated in an appropriate manner for their future location on a distributed platform. The vertical partition is of great importance because it defends the principle that those data

Serie Científica de la Universidad de las Ciencias Informáticas
Vol. 12, No. 8, Mes Agosto, 2019
ISSN: 2306-2495 | RNPS: 2343
Pág. 18-26
http://publicaciones.uci.cu

that are commonly accessed together should reside in the same fragment. Vertical partitioning is an optimization problem that has traditionally been treated with cluster analysis techniques, hierarchical and heuristic methods with two fundamental approaches: agglomerative or divisive. In recent years, the application of population meta-heuristics to this problem has increased its emphasis and focused the main research. Therefore, the objective of this article is to defend the potential of Optimization based on Dynamic Meshes as a viable meta-heuristic to solve this problem and obtain competitive results.

Keywords: *Distributed databases, vertical fragmentation, affinity measurement, optimization based on dynamic meshes.*

Introducción

El diseño de bases de datos distribuidas es un problema de optimización que implica la solución de problemáticas como la fragmentación de los datos, su ubicación y replicación.

La fragmentación es el proceso mediante el cual una relación global es descompuesta en fragmentos horizontales y/o verticales. Un fragmento vertical atiende al agrupamiento de datos en función de atributos o conjuntos de ellos, mientras que la fragmentación horizontal atiende a dicho agrupamiento en función de tuplas o conjuntos de tuplas. Típicamente, los criterios que determinan si la fragmentación y la asignación son óptimas se establecen de manera independiente, en dos pasos. En el primero se busca la "mejor" fragmentación y, en el segundo, se busca la "mejor" ubicación de los fragmentos obtenidos en el paso anterior [1].

Comparando las formas de fragmentación, la partición vertical es más complicada que la partición horizontal, debido al incremento del número de posibles alternativas [2]. También existe la fragmentación mixta.

Materiales y métodos

Este trabajo se concentra en el problema de la fragmentación o partición vertical en conglomerados sin solapamiento. Como se señala en [3][4][5], un objeto con m atributos puede ser particionado de $B(m)$ diferentes formas, donde $B(m)$ es el m-ésimo número de Bell, para m suficientemente grandes, $B(m)$ se aproxima a m^m; para m=15 este es $\approx 10^9$, para m=30 este es $\approx 10^{23}$.

Por lo tanto, es importante contar con una estrategia que reduzca de manera eficiente el número de cálculos. Aunque diferentes, dos enfoques secuenciales que conducen a agrupamientos jerárquicos parecen haber adquirido un interés particular entre los taxónomos. Una de las estrategias es el algoritmo propuesto por Ward (1963). Su idea es aglomerar los puntos o los grupos resultantes, reduciendo su número en uno en cada etapa de un procedimiento de

fusión secuencial, hasta que todos los puntos estén en un único clúster. Un algoritmo contrario ha sido propuesto por Edwards y Cavalli-Sforza (1965). La esencia de su método es la partición consecutiva de un conjunto de puntos en dos subconjuntos: primero un conjunto inicial es dividido en dos grupos, cada uno de ellos se subdivide en dos grupos más pequeños por separado, y así sucesivamente, hasta que se alcancen los puntos individuales [6].

Hoffer y Severance [7] miden la afinidad entre cada par de atributos construyendo una Matriz de Afinidad de Atributos (MAA) que sirve de base para agrupar los atributos usando el Algoritmo de Energía de Enlace (BEA, acrónimo del inglés Bond Energy Algorithm) desarrollado en [8]. La complejidad del algoritmo es razonable, del orden de $O(n)$donde n es el número de atributos. Este trabajo sirvió de motivación para la mayoría de los trabajos siguientes sobre fragmentación vertical. S. Navathe, S. Ceri, G. Wiederholdy J. Dou en [3] dividen el problema en dos etapas. Primeramente efectúan la fragmentación de las relaciones aplicando diferentes funciones objetivo empíricas que agrupan los atributos extendiendo los trabajos de Hoffer y Severance mediante el algoritmo BEA. Estos algoritmos determinan grupos de atributos en fragmentos solapados y no solapados. Posteriormente se realiza la ubicación replicada o no de fragmentos a sitios, aplicando unalgoritmo heurístico de tipo goloso (greedy). La metodología empleada es de particionamiento en lugar de agrupamiento. Cornell y Yu [9] optimizaron el trabajo de Ceri[10] desarrollando un algoritmo para la fragmentación vertical, que obtiene una partición binaria óptima para bases de datos relacionales usando información de factores físicos para disminuir el número de accesos a disco. Posteriores refinamientos son logrados aplicando un algoritmo de partición binaria iterativamente. Öszu y Valduriez[5] discuten este trabajo previo sobre fragmentación vertical en bases de datos distribuidas usando información de las frecuencias de acceso y aplican el algoritmo BEA. Los grupos de atributos son clusterizados y se usan ecuaciones de costo para definir la mejor posición a lo largo de la diagonal de la matriz clusterizada para dividir la relación enfragmentos. Muthuraj y otros autores [11] presentan un trabajo donde se argumenta que los primeros algoritmos para la fragmentación vertical son *ad hoc*, por eso se propone una función objetivo llamada Evaluador de Particiones para determinar la calidad de las particiones generadas por varios algoritmos [12].

Las meta-heurísticas basadas en población, son aquellas que emplean un conjunto de soluciones (población) en cada iteración del algoritmo, en lugar de utilizar unaúnica solución como las meta-heurísticas de trayectoria simple. Estas proporcionande forma intrínseca un mecanismo de exploración paralelo del espacio de soluciones. Dentro de esta clasificación se destacan los Algoritmos Evolutivos (Evolutionary Algorithms; EA) [13][14] y los algoritmos basados en Inteligencia Colectiva (Swarm Intelligence; SI) [15][16].Estas meta-heurísticas poblacionales son de las más estudiadas y comparten como característica fundamental que han sido inspiradas en algún proceso natural. Los EA, fueron inspirados por la teoría de la evolución de Darwin [17].Un ejemplo clásico de este tipo de algoritmos son los

Algoritmos Genéticos (Genetic Algorithms; GA) [18]. Por otra parte, los SI toman su inspiración en ejemplos biológicos de comportamiento colectivo (enjambre) como es el caso de las colonias de insectos, las bandadas de aves y los cardúmenes de peces. Dentro de estos algoritmos se encuentran; la Optimización Basada en Enjambre de Partículas (Particle Swarm Optimization; PSO) [19], y Optimización Basada en Colonia de Hormigas (AntColony Optimization; ACO) [20].[21]

En los últimos años, la aplicación de meta-heurísticas poblacionales al problema de la fragmentación vertical, ha incrementado su énfasis y centrado las principales investigaciones, con resultados muy alentadores. Por solo citar algunos ejemplos, en 2006, J. Du, R.Alhajjy K.Barker[4] tratan el problema usando Algoritmos Genéticos. Más tarde, en 2011, B. Benmessahel y M.Touahria[2] proponen una solución que usa la Optimización Basada en Enjambre de Partículas. Y más recientemente, en 2012, M.Goliy Rouhani, presentan un nuevo algoritmo de fragmentación vertical basado en Colonia de Hormigas [22].

Estudios revelan que no existe la mejor de las meta-heurísticas para solucionar cualquier problema complejo[21]. Por tal motivo, el objetivo de este artículo es defender las potencialidades de la Optimización basada en Mallas Dinámicas como meta-heurística viable para resolver el problema de la fragmentación vertical y obtener resultados competitivos.

Este trabajo presenta en la sección 2 la descripción de la metodología a seguir para aplicar la Optimización basada en Mallas Dinámicas al problema de la fragmentación vertical de bases de datos distribuidas, la descripción, características y novedades de esta meta-heurística así como los argumentos que la hacen atractiva para dicho problema. En la sección 3 se presentan las conclusiones de este artículo de posición.

Resultados y discusión

Como la inmensa mayoría de los algoritmos previos para la partición vertical de bases de datos se utilizará la Matriz de Uso de Atributos (MUA) como entrada. Esta matriz relaciona las transacciones con los atributos de la relación así como las frecuencias de acceso de cada transacción.

Hoffer y Severanceen [7] proponen el concepto de afinidad entre pares de atributos[23].

Por definición una *medida de afinidad* es una expresión matemática que permite resumir en un número el grado de relación entre dos entidades, sobre la base de la semejanza o la desigualdad entre la cualidad o la cantidad de sus atributos, o ambas [24].

Aplicando este concepto a la MUA se obtiene la Matriz de Afinidad entre atributos (MAA), que es lo que se propone en los 2 primeros pasos del Método Navathe como se indica en [25].

Sin embargo, muchos autores han criticado el uso de esta matriz. S.Chakravarthy, J.Muthuraj, R.Varadarajan y S. B. Navathe en [26] aseguran que, debido a que solo un par de atributos son involucrados, esta medida no refleja la cercanía o afinidad cuando más de dos atributos son implicados. En este trabajo se comparte el enfoque de J. Du, K.Barker y R.Alhajj, quienes fundamentan en [23] las limitaciones de la medida de afinidad como una medida de afinidad local y la necesidad de una medida de afinidad global para lograr que todos los valores de la matriz sean comparables entre sí. Se realizó un análisis de varias de las medidas de afinidad existentes y que son revisadas por A. Herrera en [24].Se decide que el Índice de Jaccard, una expresión de similitud, es una medida de afinidad global apropiada para el tema de la partición vertical de bases de datos, eliminando de su fórmula el factor que representa las ausencias conjuntas o ceros compartidos, debido a que el hecho de que en una transacción no se usen ninguno de los dos atributos del par analizado, no brinda ninguna información para el caso que ocupa.

Al aplicar la medida de afinidad global seleccionada, se obtiene la Matriz de Atracción entre Atributos (MAA*). En esta matriz simétrica, los valores quedan normalizados entre cero y uno, donde uno representa el valor máximo de similitud y cero el mínimo, por lo que todos los valores son comparables entre sí. [27].

A partir de este punto, con la información relativa a la base de datos como delas aplicaciones que acceden a la misma y la atracción entre atributos, comenzaríamos a aplicar la Optimización basada en Mallas Dinámicas (DynamicMeshOptimization, DMO).

DMO es una meta-heurística poblacional con características evolutivas donde un conjunto de nodos que representan soluciones potenciales a un problema de optimización, forman una malla (población) que dinámicamente crece y se desplaza por el espacio de búsqueda (evoluciona). Para ello, se realiza un proceso de expansión en cada ciclo, donde se generan nuevos nodos en dirección a los extremos locales (nodos de la malla con mejor calidad en distintas vecindades) y el extremo global (nodo obtenido de mejor calidad en todo el proceso desarrollado); así como a partir de los nodos fronteras de la malla. Luego se realiza un proceso de contracción de la malla, donde los mejores nodos resultantes en cada iteración son seleccionados como malla inicial para la iteración siguiente. La formulación general de la meta-heurística abarca tanto los problemas de optimización continuos como los discretos. Dicha malla se hace más "fina" en aquellas zonas que parecen ser más promisorias. Es dinámica en el sentido que la malla cambia su tamaño (cantidad de nodos) y configuración durante el proceso de búsqueda [21][28][29].

El proceso de generación de nodos en cada ciclo comprende los pasos siguientes:generación de la malla inicial, generación de nodos en dirección a los extremos locales *(nl)*, generación de nodos en dirección al extremo global *(ng)*, generación de nodos a partir de las fronteras de la malla *(nf)* y aplicación del operador de limpieza adaptativo.

El método incluye los parámetros: cantidad de nodos de la malla inicial *(Ni)*, cantidad máxima de nodos de la malla en cada ciclo *(N)*, donde $3 \cdot Ni \leq N$, tamaño de la vecindad *(k)* ycondición de parada *(M)*.

Para este caso se propone codificar nodos de la malla (las soluciones) y manipularlos según el enfoque de Cadena de Crecimiento Restringido Orientado a Grupos descrito por J. Du, R. Alhajj y K. Barker en [4][30]. También se sugiere usar como función objetivo a optimizar la ecuación propuesta como Evaluador de Particiones por Muthuraj y otros autores en [26] o la ecuación del balance entre homogeneidad interna y heterogeneidad externa propuesta por los autores del presente trabajo en [27].

A continuación se valoran algunas características y novedades de DMO que nos inclinan a defenderla como meta-heurística prometedora para tratar el problema de la fragmentación vertical de base de datos. Esta nueva meta-heurística poblacional tiene potencialidades y características similares a otros métodosexistentes (GA y PSO), que han sido aplicados exitosamente al problema en cuestión, por ejemplo:

El considerar solamente los vecinos más cercanos en la valoración de los extremos locales fue introducido en el método PSO, como se reporta en [31]; donde al determinar la mejor partícula global se toman en cuenta dos enfoques:considerar el mejor de toda la población o la mejor partícula entre los vecinos.

La atracción a zonas más promisorias del espacio de búsqueda con el acercamiento al extremo global fue introducida también en PSO; donde en su versión original [19], cada partícula es atraída por la mejor posición global delenjambre.

La concepción de expansión y contracción de la malla inicial es una forma deaumentar la población inicial con la incorporación de nuevas soluciones y luegoreducirla a través de un proceso de selección. Este elemento ha sido bienestudiado en los Algoritmos Evolutivos Generacionales [18].

La selección de la malla inicial de manera elitista también fue introducida en losAlgoritmos Genéticos como estrategia para acelerar la convergencia [21].

A pesar de compartir algunas características con otros métodos existentes, DMOincorpora otros elementos propios, como son[21]:

Utilizar en una misma iteración la atracción hacia los extremos locales de cada vecindad y el extremo global de la población. Esto representa una nueva forma derealizar intensificación y diversificación manteniendo la dirección de la búsqueda.

Se realiza una selección elitista con diversidad, en la que se tiene en cuenta tanto la calidad como la separabilidad entre soluciones al mismo tiempo. Este elemento garantiza que el método realice una profunda exploración del espacio solución,disminuyendo en gran medida el estancamiento de soluciones.

Se introduce un proceso de limpieza adaptativo, donde la distancia que garantizala separabilidad de las soluciones es un valor que decrece en función del estado del método. Permitiendo al inicio del algoritmo soluciones más distantes que alfinal de la ejecución. Este funcionamiento provoca que el método comience con un nivel alto de exploración, el cual decrece a medida que disminuye la distancia permitida con la ejecución del algoritmo. Locontrario sucede con el nivel de explotación del método.

Se incorpora una forma para guiar la exploración hacia los entornos de espacio de búsqueda a través de la generación de nuevos nodos a partir de los nodos fronteras de la malla. De esta manera se aprovecha la posición que ocupan estos nodos,explorando fuera de los entornos donde se está realizando la búsqueda.

Se probó que el algoritmo DMO presenta un alto nivel de escalabilidad, obteniéndose conel mismo resultados superiores de manera general a los demás algoritmos del estado delarte involucrados en la comparación en [21].

A pesar de tratarse de una meta-heurística bastante reciente ya ha sido aplicada a diferentes problemas, discretos y continuos, con resultados prácticos halagüeños, como por ejemplo en: aproximación de funciones continuas [21], el problema de Selección de Rasgos [32], System-LevelFault Diagnosis [29], TheFacilityLocationProblem (FLP) [33].

Conclusiones

Con el desarrollo del presente artículo de posición, se ha logrado proponer teóricamente la metodología para tratar el problema de la fragmentación vertical de base de datos usando MDO, presentando argumentos sólidos que corroboran las potencialidades de esta meta-heurística poblacional para resolver este problema NP-completo. El principio de que no existe la mejor de las meta-heurísticas para solucionarcualquier problema complejo, nos deja una puerta abierta para seguir explorando nuevas alternativas para obtener resultados competitivos y mejores, sin otra alternativa que experimentar. MDO posee características de otros métodos aplicados exitosamente a la fragmentación vertical e incorpora elementos propios y novedosos, es escalable y a pesar de ser novel ya ha sido aplicada a problemas prácticos que respaldan su efectividad.

Referencias

1. Taddei, E.,Kury, A.:Fragmentación vertical y asignación simultánea en BDD usando algoritmos genéticos, (2000),http://cursos.itam.mx/akuri/PUBLICA.CNS/2000/Fragmentaci%F3n%20Vertical%20usando%20AGs.pdf

2. Benmessahel B., Touahria M.: An improved Combinatorial Particle Swarm Optimization Algorithm to Database Vertical Partition. In: Journal of Emerging Trends in Computing and Information Sciences, vol. 2, No. 3, pp. 130--135. (2011)
3. Navathe, S., Ceri, S., Wiederhold, G., Dou, J.: Vertical Partitioning algorithms for database design. In: ACM Transactions on Database Systems, vol. 9, No. 4, pp. 680--710.(1984)
4. Du J., Alhajj R., Barker K.: Genetic algorithms based approach to database vertical partition. In: Journal of Intelligent Information Systems, vol. 26, No. 2, pp. 167--183. (2006)
5. Özsu, M.T.,Valduriez, P.: Principles of Distributed Database Systems (2nd ed.). Prentice-Hall, New Jersey, USA (1999)
6. Caliński, T., Harabasz, J.: A Dendrite Method for Cluster Analysis. In: Communications in Statistics - Theory and Methods, vol. 3, No. 1, pp. 1 -- 27. (1974)
7. Hoffer, J. A., Severance, D.G.: The Use of Cluster Analysis in Physical Database Design. In: Proceedings of the 1st International Conference on Very Large Databases. Framingham, MA, USA. (1975), http://portal.acm.org/citation.cfm?id=1282480
8. Mccormick W. T., Schweitzer P. J., White T. W.: A Problem Decomposition and Data Reorganization by a Clustering Techniques. In: Operation Research, No. 20, pp. 993--1009. (1972)
9. Cornell D. W., Yu P. S.: A Vertical Partitioning Algorithm for Relational Databases. In: Proceedings of the Third International Conference on Data Engineering. Los Angeles, CA, USA, IEEE Computer Society. (1987)
10. Ceri S. Pernici B.: DATAID-D: Methodology for Distributed Database Design, North-Holland. (1985)
11. Muthuraj J., Chakravarthy S., Varadarajan R., Navathe S. B.: A Formal Approach to the Vertical Partitioning Problem in Distributed Database Design. In: Proceedings of the 2nd International Conference on Parallel and Distributed Information Systems (PDIS 1993), Issues, Architectures, and Algorithms. San Diego, CA, USA, IEEE Computer Society. (1993)
12. Valdés L.: Asistentes para el Diseño Lógico de Bases de Datos Distribuidas, Trabajo de Diploma, Facultad de Matemática, Física y Computación, Universidad Central "Marta Abreu" de Las Villas, Cuba. (2009)
13. Bäck T., Hammel U., Schwefel H.: Evolutionary Computation: Comments on the History and Current State. In: IEEE Transactions on Evolutionary Computation, vol 1, No. 1, pp. 3--17. (1997)
14. Cano J.R., Herrera F., Lozano M.: Using Evolutionary Algorithms as Instance Selection for Data Reduction in KDD: An Experimental Study. In: IEEE Transactions of Evolutionary Computation, No. 7, pp. 561--575. (2003)

15. Bonabeau E.: Swarm Intelligence: From natural to artificial systems. 1ra ed., Oxford University Press, USA. (1999)
16. Engelbrecht A. P.: Fundamentals of Computational Swarm Intelligence. John Wiley and Sons. (2006)
17. Darwin C.: On the Origin of Species. John Murray, London. (1859)
18. Goldberg D. E.: Genetic Algorithms in Search. Optimization and Machine Learning. Addison-Wesley Publishing Company: University of Alabama. (1998)
19. Kennedy J., Eberhart R. C.: Particle swarm optimization. In: IEEE International Conference on Neural Networks. Piscataway, New York, USA, vol. 4, pp. 1942--1948. (1995)
20. Dorigo M.: Optimization, Learning and Natural Algorithms. Phd. Dipartimento di Elettronica. Politecnico di Milano. (1992)
21. Puris A. Y.: Desarrollo de meta-heurísticas poblacionales para la solución de problemas complejos, Tesis en opción al título de Doctor en Ciencias Técnicas, Especialidad Informática, Centro de Estudios de Informática, Departamento de Ciencia de la Computación, Facultad de Matemática, Física y Computación, UCLV, Cuba. (2009)
22. Goli M., Rouhani R. M. T.: A new vertical fragmentation algorithm based on ant collective behavior in distributed database systems. In: Knowledge and Information Systems, vol. 30, No. 2, pp. 435--455. (2012)
23. Du, J., Barker, K., Alhajj, R.: Attraction-A global affinity measure for data base vertical partitioning, (2003), http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/barker:ken.html
24. Herrera, A.: La clasificación numérica y su aplicación en la ecología. Sammenycar C. x A., Santo Domingo, República Dominicana (2000)