Mass Data Query Optimization Based on Multi-objective Co-evolutionary Algorithm

Zhang Ting[1]

1. Information Technology Center, Beijing Jiaotong University, Beijing, 100044, China

Key words: mass data; query optimization; evolutionary algorithm; multi-objective; cooperative computing

Abstract. multi-connection database query optimization belongs to a kind of typical complex problem and cost of optimal query strategy obtained from traditional Particle Swarm Optimization Algorithm is relatively high under some conditions and it is easy to fall into local optimal solution. Based on Quantum Particle Swarm Optimization Algorithm, the paper puts forward a kind of optimal algorithm for database query, namely, mass data query algorithm based on multi-objective co-evolutionary algorithm to improve optimization efficiency of database query and optimize performance of algorithm of the paper in solution of database query optimization problems by simulation experiment. The paper puts forward a kind of Gaussian Mutation Quantum Particle Swarm Optimization Algorithm and introduces Gaussian mutation to avoid prematurity phenomenon. Experimental result shows that algorithm of the paper can obtain more optimized query effect when solving multi-list connection database query optimization problems.

Introduction

With continuous development of database technology, query optimization is increasingly significant and purpose of query optimization is to translate query requirement of users into an effective query processing strategy to reduce loss caused by blind query [1].

Aiming at database query optimization problems, domestic and overseas scholars make many efforts from different perspectives and put forward many optimized algorithm. The most primitive database query optimization algorithm adopts exhaustion algorithm or other morphing algorithm; when the number of connection relation is relatively small, it can get favorable query effect, while query connection number of modern database management system is generally large and this kind of algorithm is useless [2]. Later, scholars put forward that using dynamic planning algorithm to solve database query optimization problem and its query efficiency is hard to be accepted [3]. Because database query optimization problem is a multi-constraint combination and optimization problem and is a NP difficult problem, some scholars introduce heuristic algorithm into database query. Shuai Xunbo proposed a kind of query optimization method based on genetic algorithm; defines firstly a new query implementation plan cost model and carries out query optimization based on the cost model and making use of genetic algorithm [4]. T.V. Vijay Kumar puts forward query strategy to carry out query optimization and find out least nodes number by utilizing correlation degree of query data of genetic algorithm [4].

Mass data query algorithm based on multi-objective co-evolutionary algorithm

Particle Swarm Optimization (PSO) Algorithm

Assumed that xi=(x1,x2…,xd,…,xD) represents position vector of the ith particle; where, i=(1,2,…,m); vi=(vi1,vi2,…,vid,…,viD) is flight speed vector of particle I; history of the ith particle should be represented by Pi=(Pi1,Pi2,…,Pid,…,PiD); the best point that all particles in swarm pass is represented as Pg=(Pg1,Pg2,…,Pgd,…,PgD).

In iteration of each time, position and speed of particle shall be updated according to the following equation:

$$v_{id}^{k+1} = v_{id}^{k} + c_1 r_1 \left(P_{id} - x_{id}^{k} \right) + c_2 r_2 \left(P_{gd} - x_{id}^{k} \right) \qquad (1)$$

$$x_{id}^{k+1} = x_{id}^{k} + v_{id}^{k+1} \quad (2)$$

Where, i=1, 2,…, m, d=1, 2, …, D; wmax is initial weight; wmin is final weight; itermax is maximal number of iterations; k is present number of iterations; c1 and c2 are learning factor; r1 and r2 are random numbers that distribute evenly.

Quantum Particle Swarm Optimization (QPSO) Algorithm

In search process of PSO Algorithm, search of particle cannot cover the whole feasible space and it cannot be ensured that globally optimal solution can be found. In quantum space, because particle meets nature of state of aggregation, it can be searched in the whole feasible space. Therefore, after researching convergence behavior of particle, Sun puts forward Quantum Particle Swarm Algorithm (QPSO).

Assumed that search space is D-dimension space, there are n particles in swarm. QPSO Algorithm updates position of particle i by the following formulas:

$$\begin{cases} x_i(t+1) = p - \alpha \times \left| mbest - x_i(t) \right| \times \ln \dfrac{1}{u} & if\ (u > 0.5) \\ x_i(t+1) = p + \alpha \times \left| mbest - x_i(t) \right| \times \ln \dfrac{1}{u} & f\ (u \le 0.5) \end{cases} \quad (3)$$

Where, u is random number that distributes on (0, 1) evenly; p is defined as a random number between extreme value of particle individual pbest and overall extreme value of present swarm gbest, namely:

$$p = \frac{pbest \times b_1 + gbest \times b_2}{b_1 + b_2} \quad (4)$$

Where, b_1 and b_2 are random numbers on (0, 1).

α is contraction-expansion coefficient and contraction speed of algorithm can be controlled by adjusting α, namely:

$$\alpha = 0.5 + \frac{0.5 \times (N - t)}{N} \quad (5)$$

Where, N is maximal number of iterations.

Besides, mbest in formula (6) is defined as mean value of individual extreme value in the whole particle swarm, namely:

$$mbest = \frac{1}{m}\sum_{i=1}^{n} pbest_i = \left(\frac{1}{m}\sum_{i=1}^{n} pbest_{i1}, \frac{1}{m}\sum_{i=1}^{n} pbest_{i2}, \sum_{i=1}^{n}\frac{1}{m}\sum_{i=1}^{n} pbest_{iD}\right) \quad (6)$$

Quantum Particle Swarm Optimization (algorithm of the paper) Algorithm with Gaussian Mutation

QPSO Algorithm is the same with other swarm intelligence algorithm and will be confronted with problem of swarm diversity loss, which will cause that particle will get together gradually and get into locally optimal solution. Mutation operator can be used to prevent swarm diversity loss and permit search space of larger area; therefore, some scholars introduce mutation operator into QPSO Algorithm and change present particle to avoid falling into local optimal solution by search stagnation. At present, QPSO researches about mutation operator are relatively few; where, Literature [1] adopts Cauchy mutation operator to change average optimal position and overall optimal position:

$$\text{mutate}(mbest) = mbest + \phi\delta \quad (7)$$

$$\text{mutate}(P_g) = P_g + \phi\delta \quad (8)$$

Where, ϕ is mutation size and δ is random variable of Cauchy distribution and probability density function is:

$$g(x) = \frac{a}{\pi(x^2 + a^2)} \quad (9)$$

Where, a is a scale parameter and it decides shape of distribution and it is controlled by annealing function:

$$a = a_0 (CR)^k \tag{10}$$

In the paper, the paper adopts mutation operator of Gaussian distribution to substitute Cauchy distribution and applies it to QPSO Algorithm. Advantages of Gaussian distribution are probability of it for small disturbance to position is relatively large and probability density function is:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \tag{11}$$

Where, variance σ^2 controls width of distribution.

It is observed from position update formula (8) that local optimum can be avoided by applying mutation operation into average optimal position mbest(k) and finding a new mbest(k).

Adopt 4 standard test functions: Sphere, Griewank, Ackley and Rastrigin to test performance of QPSO Algorithm, Cauchy mutation operator and QPSO (G-QPSO) Algorithm and algorithm of the paper. Images of 4 test functions with dimension of 2 are shown as Fig. 1.

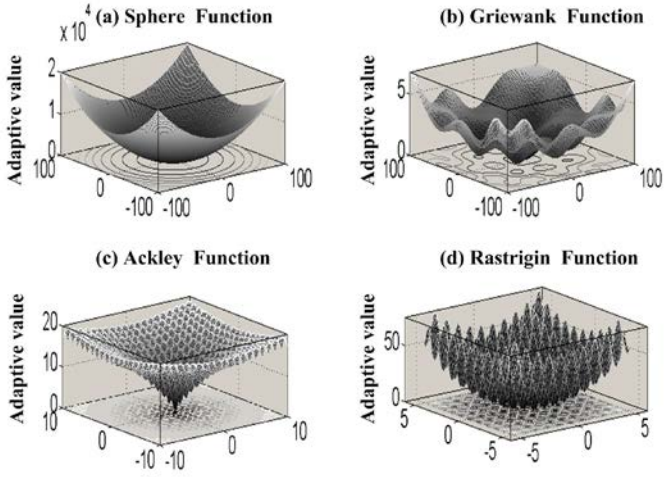
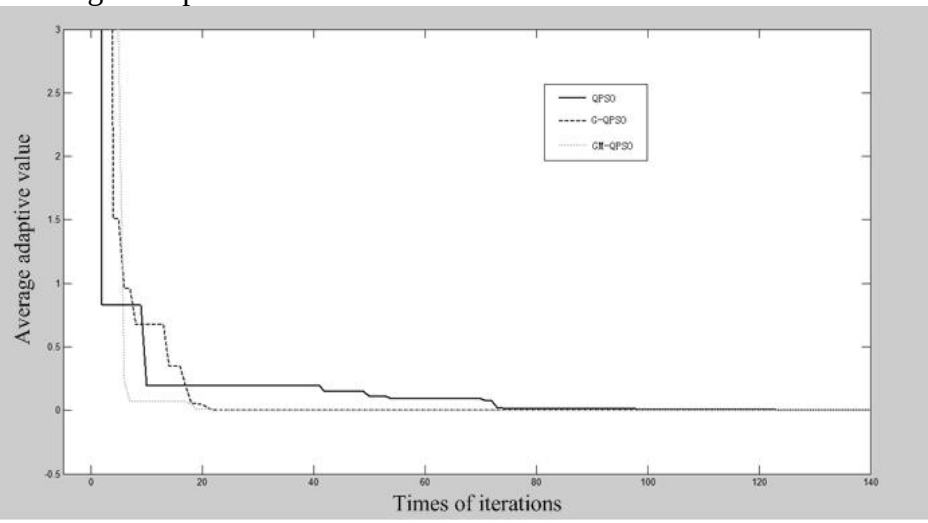
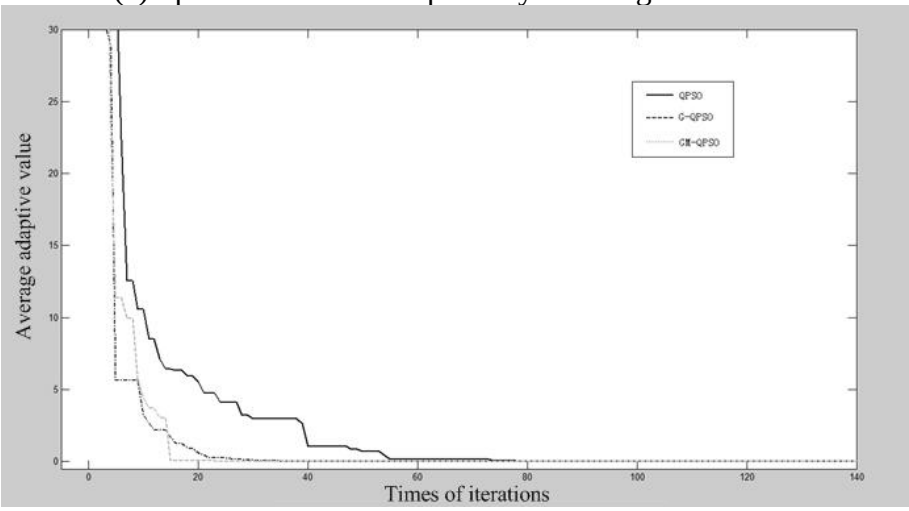


Fig.1 Images of 4 Test Functions with Dimension of 2

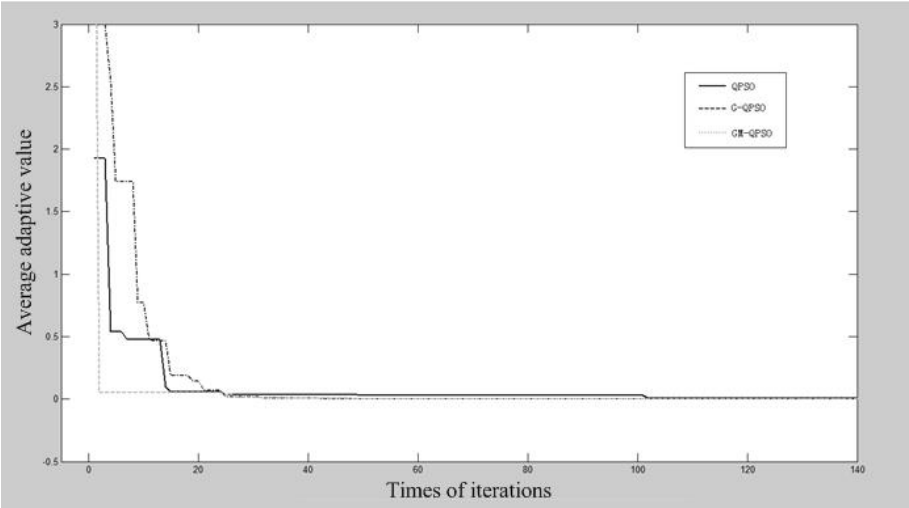
It can be known from Fig.1 that Sphere function only has one overall optimal value and it is easy to be solved; Griewank function is not smooth and continuous near overall optimal value, which has increased search difficulty; Ackle function only has one overall optimal value, while it almost presents discontinuous state near overall optimal value and search difficulty is very large near overall optimal value. Rastrigin is a kind of pathological function and overall optimal value has very deep slit with local optimal value and many search algorithms can hardly find overall optimal value and operation results of all algorithms are shown as Fig.2. It can be known from Fig.2 that for all functions, convergence speed of algorithm of the paper is obviously superior to contrast algorithm. It means that algorithm is the paper obtains superior overall search ability, convergence precision and convergence speed.



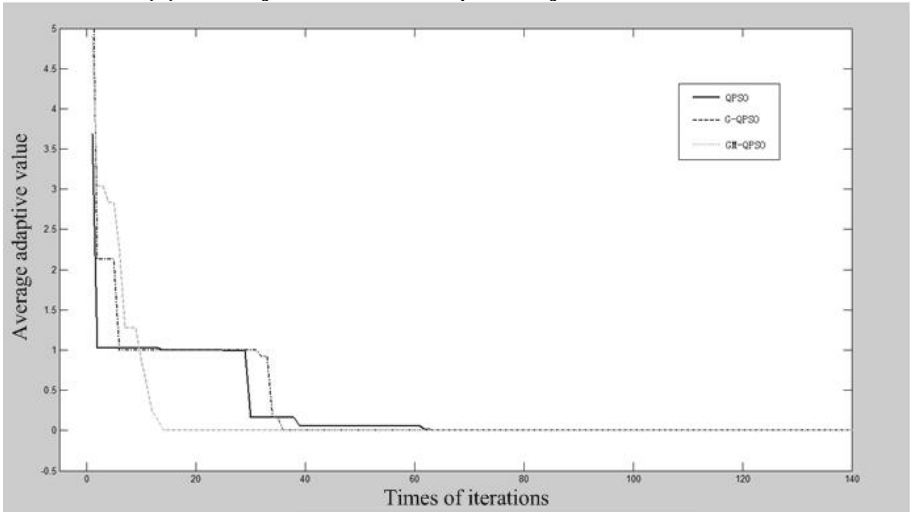
(a) Sphere Function Adaptability Convergence Curve



(b) Griewank Function Adaptability Evolution Curve



(c) Ackley Function Adaptability Evolution Curve



(d) Rastrigin Function Adaptability Evolution Curve

Fig. 2 Convergence Performance Contrasts of QPSO, G-QPSO and Algorithm of the Paper

Experiment

Result and analysis

After experiments for several times and solution for mean value, change curve for cost ratio for searching optimal query scheme by QPSO, G-QPSO and algorithm of the paper is shown as Fig. 4.

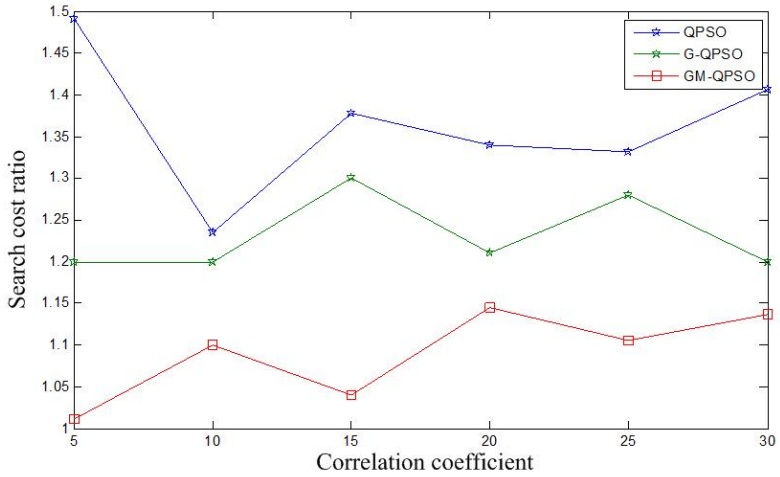


Fig.3 Search Cost Consumption Ratio of Query Optimization Algorithm of Various Database

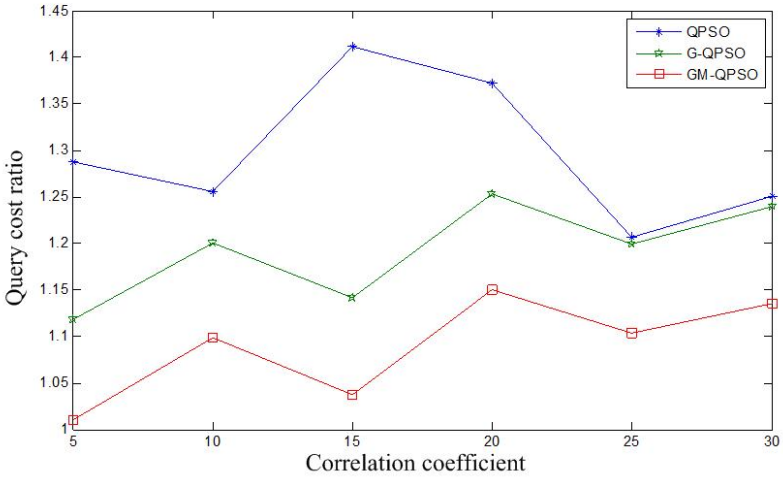


Fig.4 Query Cost Consumption Ratio of Query Optimization Algorithm of Various Database

It can be concluded after analyzing results in Fig. 3 and Fig. 4:

(1) When connection numbers of database query are relatively few (less than 10), algorithm query and implementation cost of QPSO, G-QPSO and algorithm of the paper have no difference on performance after contrast.

(2) With increase of numbers of connection relations, algorithm query and implementation cost of QPSO, G-QPSO and algorithm of the paper all increase gradually; when connection number of database query is relatively numerous (larger than 20), advantages of G-QPSO and algorithm of the paper will embody and it means that G-QPSO and algorithm of the paper can meet the requirement of query and optimization for management system of large-scale database.

(3) Compared with G-QPSO algorithm, advantages of algorithm of the paper are relatively obvious because Gaussian mutation operator has increased swarm diversity and has effectively realized balance of overall and local search of algorithm and it not only has improved database query efficiency, but also has reduced implementation cost so that optimal database query plan can be found in quicker speed.

Evaluate quality of convergence condition and optimal query scheme of optimal query scheme for algorithm search to further analyze query optimization effect of algorithm of the paper to data connection and the result is shown as Table 1.

Number of relations	Number of quantum and particle	Number of iteration of optimal query scheme			Implementation time of optimal query scheme		
		QPSO	G-QPSO	Algorithm of the paper	QPSO	G-QPSO	Algorithm of the paper
5	4	170	158	135	52.46	44.54	37.17
10	5	180	167	140	97.58	84.97	61.56
20	10	475	368	245	266.74	209.27	183.05
40	15	490	440	332	401.21	304.91	235.61
60	20	480	445	362	354.35	304.86	265.56

In table 1, when correlation coefficient n=20, the number of quantum and particle is m=10; optimal database query scheme can be obtained after 475 times iterations of QPSO and its implementation time is 266.74s; optimal database query scheme can be obtained after 368 times iterations of G-QPSO and its implementation time is 209.27s; convergence speed and quality of optimal solution can be improved significantly. Database query scheme obtained after times iterations of algorithm of the paper is superior to QPSO algorithm and G-QPSO algorithm and its implementation time is 183.05s. It can be got after analyzing data in Table 1 comprehensively: convergence speed for solution accelerates and quality of optima solution is improved greatly of G-QPSO algorithm compared with QPSO algorithm; while, search performance of algorithm of the paper is improved significantly so that premature convergence phenomenon can be better avoided and superior query scheme obtained with quicker speed convergence has improved database query efficiency to some degree.

Conclusion

The paper firstly introduces mutation operator in genetic algorithm into Quantum Particle Swarm Optimization Algorithm so that particle position moves within small range of quasi-optimal solution to improve overall search ability of algorithm and then applies it to database query optimization problem for solution and tests performance of algorithm of the paper by simulation experiment. It is shown from results that algorithm of the paper accelerates convergence speed for database query optimization solution and obtains query optimization scheme with higher quality.

Reference

[1] Bethencourt J, Sahai A, Waters B. Ciphertext-Policy Attribute-Based Encryption[C]// IEEE Symposium on Security and Privacy. IEEE Computer Society, 2007:321-334.

[2] Lewis D D, Yang Y, Rose T G, et al. RCV1: A New Benchmark Collection for Text Categorization Research[J]. Journal of Machine Learning Research, 2004, 5(2):361-397.

[3] Zhai C, Lafferty J. A study of smoothing methods for language models applied to Ad Hoc information retrieval[C]// International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2004:179-214.

[4] Ramakrishnan N, Kumar D, Mishra B, et al. Turning CARTwheels: an alternating algorithm for mining redescriptions[C]// Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2004:266--275.

[5] Elkin M D P L. UMLS Concept Indexing for Production Databases: A Feasibility Study[J]. Journal of the American Medical Informatics Association, 2001, 8(1):80-91.