TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 9, September 2014, pp. 6921 ~ 6931
DOI: 10.11591/telkomnika.v12i9.4355

The Hybrid Probabilistic Query Algorithms Based on Inconsistent Database

Gao Hongyu
Shandong Wan Jie medical college, Shandong Zibo, China
E-mail: 35113479@qq.com

Abstract

Integrity constraint is important to make data certain in relation database. Though there is plenty of uncertain information that is valuable and need to be searched and to be used. Combined with probabilistic database theory and on the basis of summarizing former results, this paper gives a new query plan aiming at inconsistent database. It uses the constraint methods including union, product, subtraction, selection, projection and link to repair inconsistent data effectively. Its probabilistic calculation with four elements and probabilistic query rewriting can overcome shortcomings of inconsistent databases. The experiments show these methods can decrease conflict of data.

Keywords: *inconsistent databases, probabilistic data model, data clean, query rewriting, query of constraints*

1. Introduction

As is known to all, the magic weapon that a relational database rules the database field enduring is descriptive query language SQL and its basic relational algebra operation. Relational algebra is one of the characteristics of clear query semantics, another is used to support query optimization, and its advantage comes from the simple and clear data model - the relationship, which has a sound mathematical basis and system transformation rules, and with uncertainty information as the research object, with integrity constraints, such as the entity integrity, domain integrity, referential integrity, and user-defined integrity, these constraints effectively to ensure the integrity and effectiveness of the data in a relational database. The data conform to the real world entity rules, in favor of the formation and query of data.

Non-uniform database, in a nutshell is the transgression of the database data integrity constraint, which contain violate integrity constraints in data and uncertain data. The real world there is many don't comply with the requirements for data integrity, the content of uncertain information, such as market prediction, fuzzy analysis and so on. It will need to establish a can handle data integrity, the content database model uncertainty, in terms of dealing with uncertain information, have put forward a variety of probabilistic relation database model to deal with uncertainty information. Using probability theory [1] proposed relational schema extensions, using a probability of 1NF relational view. Using probability theory, semantic redefines the projection operation selection and connection, is one of the earliest put forward extension probability relations literature, but incomplete scaling problem; Literature [2] gives a limited to the sum of all the probability of tuples probability relationship model which equal to 1, but case is not complete consideration of the sum of probability is less than 1. Literatures [3, 4] proposes a probability data for relational database model, to further improve the probability of the sum is less than 1, which put forward four tuples (that is, the probability of object properties, static properties, dynamic properties, attributes) to solve the key attributes of the static and dynamic properties of probability problems, which greatly expand the relational database's capabilities to deal with uncertain information, but little consideration to the inconsistency data; Literatures [5, 6] mainly defines the basic concept of database consistency and non-uniform data query some methods, such as data cleaning, thinking the consistency of database management is a strategy data cleaning, identify and correct errors in the data and restore the database to a consistent state, and puts forward the main methods to deal with inconsistent data such as consistent query, modify, query rewriting and so on, but these information is the main research

Received September 8, 2013; Revised Apr 24, 2014; Accepted May 17, 2014

object or deterministic data, do not apply to the uncertain data. Literatures [7-9] from the angle of probability and statistics to deal with uncertain information, and puts forward the concept of complete definition and the corresponding query rewrite the relationship. Considering the operation of the object mainly is the static property, the operating object of another attribute is only made a simple prompt dynamic property. Literatures [10, 11] from XML semi-structured angle discuss the probability of the query method of non-uniformity of database operations which have important reference value.

But with the rapid development of Internet, the network has quickly become an important means of information dissemination and exchange. Especially on the Web, has a very rich source of data, most of these data is change, some of them are regular changes, some of them are irregular changes, or changes in the law is uncertain. Literatures [12, 13] For regular change can take advantage of the relational database, and for a large number of irregular data is unable to use SQL queries The accurate data for the production of large numbers of these irregular does not always maintain the consistency of a relational database, data conflicts tend to, but they still have use value of information, not simply discarded. But some information is not accurate, it violates the principle of relationship database consistency, therefore, to find out an effective method to the consistency of inconsistent information into information, and can perform the corresponding accurate query becomes the key to solve the problem.

Particle swarm optimization and ant colony algorithm as the current group of intelligence research fields in the two main algorithms, rely on is a probability search algorithm/ As a kind of evolutionary computation method, particle swarm optimization (PSO) algorithm is to simulate the behavior of the birds feed on, based on a series of iterative optimization algorithm, the first initialized to a group of random solutions, through the iterative search for the optimal solution, has the following characteristics: with a multi-point search capabilities, to establish balance between diversity and centralized search. And ant colony algorithm is also in recent years, the birth of the stochastic optimization method, mainly through the information transmission between ant colony and achieve the goal of optimization and its advantages: one is to have positive feedback mechanism, through the pheromone update the query optimal path; 2 it is general-purpose stochastic optimization method, it is not a simple simulation of real ants, but is more integrated into people’s intelligence; Three is a global optimization method, not only solve the single objective optimization problem, and can solve the multi-objective optimization problem. So combining with the probability problem, we conduct exact solutions using PSO and ant colony algorithm. Looking forward to get win-win in terms of optimization performance is to do complementary advantages in time performance.

Compared with other query optimization algorithm, which has the following advantages: 1) Because individual fault affects to the whole problem solving, without centralized control constraints do not ensure that the system has stronger robustness; 2) In a direct way of communication to ensure the scalability of the system, due to the increase of the individuals in the system and increase the communication overhead is less; 3) Three parallel distributed arithmetic model, can fully use multiple processors, such distributed more suitable for the working state of the network environment; 4) The continuity of 4 for problem definition without special requirements; 5) ability of each individual in the system is very simple, every individual execution time is short, and the algorithm implementation is easy.

In this paper, the uncertainty of inconsistent database information is proposed to the deterministic algorithm. Query in this article about the probability and the consistency of database operations research as the foundation, it is by increasing the probability constraint conditions. The accurate extraction of data element characteristics, expand the T-SQL relational expression’s ability to process data, not only can make the uncertain information into certain information, and adopt the method of probabilistic query to query rewrite of inconsistency data, helping the promptness and accuracy of the query.

2. Problem Descriptions

2.1. Problem Definition

Definition 1 (Consistency): A pattern å R has a set of database integrity constraints on the set in the literature [5]. If the pattern of a database instance on the R and R satisfy å , which called instance r is consistency, remember to $r \models \Sigma$; Otherwise is consistent, remember to $r \nvDash \Sigma$.

Definition 2 (the inconsistent database): r is an example of the database D, for any integrity constraints on the database D (IC), if $\exists r \models IC$, D is consistency database, otherwise is consistent.

Definition 3 (short, the difference between): There are two database instance r and r' on a model of R . Given a set of database query constraint set Σ , database on the Σ is consistent. The difference between R and R' is:

$$\Delta(r,r') = (r \otimes r') = (r - r') \cup (r' - r) \neq null$$

Definition 4 (query integrity constraint): Entity rules corresponding to the pattern cannot be created on the R integrity constraints Σ . A given query Q must satisfy the Σ on R , says $R \models QIC$, otherwise $R \nvDash QIC$. Real world entities corresponding to the model, as may be inconsistent data integrity constraints in violation of the patterns, so cannot be created, but in the query results must satisfy these constraints that given, with all the tuples satisfy the query selection conditions, QIC to get consistent results again set to filter the results [14].

Defined 5 (repair) There is a set of integrity constraints on a pattern R collection Σ . r is an instance on the R . R' is R a subset meet of Σ in R . If R' satisfy the following conditions:

1) $R' \models \Sigma$

2) There is no another instance $r' \models \Sigma$, and meet the $\Delta(r,r') \subset \Delta(r',r)$. Suggests that R' is a meet the largest subset of Σ on the model R , called R' is a repair of R .

Definition 6 (query rewrite): R is a relation. Σ is a set of constraints that I is an instance on the R . q is a query Boolean value on the R . For Σ in every instance I have $\Sigma(q,I) = true$, says I satisfy the consistency on R . The query value q satisfy the query conditions, otherwise the $\Sigma(q,I) = false$. I does not meet the consistency on the R . q query failed in values, namely, given a SQL query q and a set of key constraints Σ , will rewrite the SQL query q into another Q_c . Q_c can return consistent results. The rewrite is independent of the data and work for every inconsistent database.

Such as,

$$\Psi(x_1, x_2, ..., x_k) = select * from\, T$$
$$where\, f(x_1, x_2, ..., x_k)$$

Among, $f(x_1, x_2, ..., x_k) = R \wedge (A_1 = x_1) \wedge ... (A_k = x_k)$.

Definition 7 (consistent query): R is a given model, a set of IC set Σ and a database instance on the r . r can be consistent, in violation of Σ some integrity constraints in the IC , a repair R' is a subset of the R . At R , R' meet Σ in IC , the distance R' and I must be minimized. Repair is not the only commonly, each repair for the consistency of database (IDB) may be meet Σ . An example of a given r , for every repair on the R' . A query Q meet on r on $r \models Q[t]$, the tuple t is consistent.

So a tuple on the probabilistic relational schema is into the key properties, static properties, dynamic properties, probability. Its description is no longer a simple object in traditional relational data model, but a dynamic object (including dynamic object itself and the dynamic properties) of an event with the probability of the event to occur. Formally still expressed with two-dimensional table, it essentially describes probability of a certain event occurs, the different line describes a different event and probability, are not allowed to appear on the same events in the same relationship [15-16].

2.2. Formal Presentation

As shown in Table 1 and Table 2, if you want to find the network course grade of name Bo Liu, or find online courses whose score is 97 points, is unable to provide accurate information, because Bo Liu network curriculum may is 97, it is also possible is 62, or the network course grade may be Bo Liu 97, which is also may be WenMu.

Table 1. Student Selection Course List

	Student id	Name	Courses	Probability
s1	04331008	Boliu	Network	0.8
s2	04331009	Yangmu	English	0.6
s3	04331008	Boliu	Operating system	0.2
s4	04331007	Wenmu	Network	0.4
	…	…	…	

Table 2. Cause Result Table

	Course code	Course	Grade	Probability
t1	10028	Network	97	0.9
t2	10016	English	85	0.7
t3	10028	Network	62	0.1
t4	10014	English	58	0.2
	…	…	…	

Table 1 Student id is as the key attributes, name is a static property, course is dynamic properties. The probability of 0.8 said the probability of student id 04331008 selection network course is 0.8, Figure 1 is the use of concise tree to show this relationship according to Table 1 and 2.

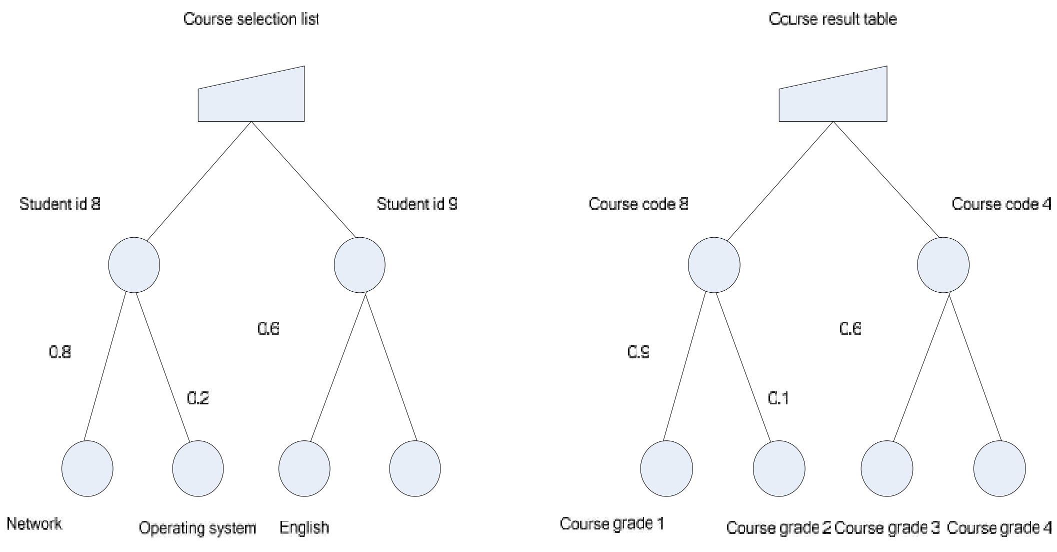


Figure 1. Course Selection and Score Probability Tree

Table 3. Probability Table of English Grade

Course code	Course	Grade	Probability
10016	English	85	0.7
10016	English	58	0.2
10016	English	null	0.1

3. Methods of Probabilistic Query and Algorithms
3.1. Particle Swarm Optimization (PSO) Algorithm

(1) Particle Swarm Optimization

Particle swarm optimization algorithm (PSO) is a kind of evolutionary computation technology comes from the study of birds feed on behavior and invention. First put forward by Dr Barnhart and Dr Kennedy in 1995 [17]. In the process of birds feed on, every bird can be regard as a particle, every bird, according to two factors determine the feeding path itself before a bird is taken by the optimal path and the other one is the optimal path through the other companion [18].

Mathematically, assume that the search space is d dimension, which $X_i^t = (x_{i1}^t, x_{i2}^t, ..., x_{id}^t)$ is the first particle in the dimension of D vector. X_i^t can be seen as a potentially viable solutions. By the changes of the particle velocity, the optimal solution is realized by the dimension D bad to search. The first of the particle's velocity is defined as: $V_i^t = (v_{i1}^t, v_{i2}^t, ..., v_{id}^t)$. $P_i^t = (p_{i1}^t, p_{i2}^t, ..., p_{id}^t)$ represented particles within the number of iterations to obtain the optimal value in iterations, said all the particles within the number of iterations to obtain the optimal value. Practice through the optimization value determined by the problem of fitness value to evaluate the particle from "bad" to "good". Each particle by the above two extreme value constantly update themselves, so as to produce a new generation of groups [19].

Particles i will be subject to the following formula to update their speed and position:

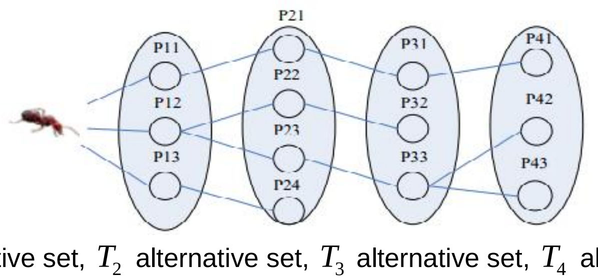
$$v_{id}^t = wv_{id}^{t-1} + c_1 rand()(p_{id}^t - x_{id}^t) + c_2 Rand()(p_{gd}^t - x_{gd}^t) \quad (1)$$

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

Which c_1, c_2 are constant and are called learning factor; $rand()$ and $Rand()$ are the random Numbers on [0, 1], w is for the inertia weight (inertia weight). Formula is composed of three parts, part 1 that particle velocity, illustrates the particle current status; Part 2 is the cognitive part (which the modal), mean the particles themselves to think; Part 3 is the society part (social modal). Three parts together determines the space search ability of particle. Part 1 has the ability to balance the global and local search. Part 2 make the particles have enough strong global search ability, avoid the local minimum. Part 3 shows the information sharing between particles. In the three parts of the particles together determine the effectively reach the best position under the action [20].

In addition, the particles constantly adjust their position according to speed, but also is limited by a maximum speed $V_{\max}$. When V_i more than $V_{\max}$, it will be limited to $V_{\max}$.

3.1. Ant Colony Algorithm



T_1 alternative set, T_2 alternative set, T_3 alternative set, T_4 alternative set

Figure 2. The Optimization Selection Probability of the XML Node Resources Based on Ant Colony Algorithm

Using ant colony algorithm to solve the XML query optimization selection problem, which is shown in Figure 2. In the figure, ovals represent task T_i corresponding alternative resource set, solid round represent alternative path corresponding to the node. To solve problems from each alternative resources concentration options out of a resource node, make all choose the path of minimum cost, get the results of the optimization. If each node resources as ants climb over a node, the problem solving process can be transformed to how to according to the $T_1, T_2, ..., T_n$ order to select a path, which makes the ants climb over the path through the largest total probability, and compared the number of minimum. In Figure 2, in order to facilitate the ant search resource node, uniform number has carried on the all the resource node. 0 is for the virtual resource node, as the starting point of the ants crawling [21-22].

In order to make the problem suitable to be solved by the ant colony algorithm, still need through appropriate transformation, due to the application of ant colony algorithm to solve the problem is the key to determine the change of the pheromone, therefore the following two type of pheromone updates are used:

$$\Delta\tau_p^k = Q_1 / L_p^k \tag{3}$$

$$\Delta\tau_g^k = Q_2 / L_g^k \tag{4}$$

In Equation (3), $\Delta\tau_p^k$ for local pheromone updating, which represents the ant k every climb over a distance, each choosing a resource node, local pheromone is left in this way. Q_1 is constant, represents the strength of the pheromone. L_p^k as the journey length, i, j, respectively for the starting point and end point of the journey, L_p^k confirmed by with Equation (5).

$$L_p^k(i,j) = \begin{cases} C_j + trc_{ij} & (i,j) \in A \\ C_j & 0.w \end{cases} \tag{5}$$

With C_j for the cost of resource node j, trc_{ij} are order constraints for the operation cost of resource nodes i, j . A is the set of directed edges in the graph constraint for task order. In Equation (4), $\Delta\tau_g^k$ as a global pheromone updating, represent the strength of the pheromone, L_g^k for (compare number) is the total length of the path with $allowed_k(t)$, which represent the first t ant allowed to crawl resource nodes. With an array of $tabu_k$ ant k has climbed resource nodes, such as $allowed_k(0) = \{1,2,3\}$, $allowed_k(1) = \{4,7\}$, $tabu_k(0) = 0$, L_g^k can be calculated by the following formula:

$$L_g^k = \sum_{t=0}^{n-1} Lp(tabu_k(t), tabu_k(t+1)) \tag{6}$$

3.2. Probabilistic Query and Algorithm Optimization

According to the characteristics of the XML tree, must carry on the probability inquiry. First to determine the probability of each node and the determination of probability need combined with the specific operation. If adopts the model of tree search, the process is not simple, need to be point by point comparison. Using ant colony algorithm, the prophase of probability processing speed is slow, and the particle swarm algorithm because do not need too many constraints, so it's suitable for a random query in the earlier time [23].

1) First conduct formalized description to the whole operation. Suppose, there has n data node set $N = (1, 2, ...n)$. The nodes have order relation before and after the node. P_j is

the direct antecedent nodes for activity j set in the node j before it, any sequence activity $P_{ii} \in j$, cannot conduct before didn't finish the former; Each node can be executed multiple times, the time needed for execution to the corresponding node j for dj a time unit. A set of query data for a given K, how to ensure that the query within given time T, and the time spent the shortest, otherwise fail.

First introduced two decision variables:

$$x_{jt} = \begin{cases} 1 & Node\ j\ in\ time\ t\ is\ proceeding \\ 0 & Others \end{cases}$$

The implementation time of the node j:

$$ET_j = \min_{1 \le t \le T} \{tx_{jt} \mid x_{ji} = 1\}$$

The end execution time node j :

$$FT_j = ET_j + DJ$$

Mission K based on query time T limited mathematical model description is as follows:

The objective function:

$$\min \max_{1 \le j \le n, 1 \le t \le T} (tx_{jt}) \quad (7)$$

$$\min \sigma^2 = \sum_{k=1}^{K} \frac{1}{P_k} \frac{1}{T} \left(\sum_{t=1}^{T} \left(\sum_{j=1}^{n} tx_{jt} - \overline{T_k}\right)^2\right) \quad (8)$$

Among them:

$$\overline{T_k} = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{n} tx_{jt} \quad (9)$$

$T = \min FT_K + d_K, P_k$ For selecting nodes probability.

Constraints:

$$\sum_{t=1}^{T} x_{jt} = d_j,\ j = 1, 2, \ldots n \quad (10)$$

$$FT_j - FT_i \ge d_j \quad \forall i \in P_j,\ j = 1, 2, \ldots n \quad (11)$$

$$\sum_{j=1}^{n} tr_{jk} \le T_k^t \quad k = 1, 2, \ldots, K;\ t = 1, 2, \ldots T \quad (12)$$

$$xji = 0\ or\ 1,\ j = 1, 2, \ldots, n;\ t = 1, 2, \ldots, T \quad (13)$$

Objective function (7) represents the query activity time as short as possible and feasible solutions; Target (8) represents the minimum variance according to the actual probability of time; Objective function (9) represents the average requirements value under the condition of query activity time is the shortest and each node within the scope of the query time;

Constraint (10) represents no matter how to arrange each nodes, its execution time is fixed; Constraints (11) means before each query process must meet tight relationship; Constraint (12) means at the end of each session using various time amount cannot be greater than the amount of time spent. Constraint (13) is the variable constraints.

2) Algorithm solve

(1) First, a series of random numbers generated by random function sequence according to the knot points, and then corresponding sequence according to the size of scheduled again. In the series of sequences, the sequence selected to meet the requirements set for the corresponding solving particles, which method is listed the initial group of X and the direction of the initial group of V then use particle swarm optimization (PSO) algorithm, starting from the initial group. Particles according to their own and partner's flight experience constantly adjust the position and speed, so as to produce a new generation of groups.

$$V_i^{k+1} = wV_i^k + c_1 Rand()(PB_i^k - X_i^k) + c_2 Rand()(GB^k - X_i^k) \tag{14}$$

$$X_i^{k+1} = X_i^k + V_i^k \tag{15}$$

In the equation, V_i^k is i particle speed in the k iteration; X_i^k is for particle i place in the first k iteration; PB_i^k For particle i individual extreme value; GB^k as global extreme value; $Rand()$ for random Numbers on the interval [0, 1]; w is inertia weight; c_1 is the cognitive factor, adjust to PB_i^k flight step length; c_2 is social factor, adjust to GB^k step length of the flight The iteration process, the particle velocity and position are restricted to a particular range, at the same time, the PB_i^k and GB^k constantly updated, the final output of the GB^k is global optimal solution.

(2) to find success in the node, using the ant colony algorithm rules corresponding to the probability of nodes can be calculated using the following formula pheromone on the path to the next should choose path probability.

$$\tau_{ij}(t+n) = (1-\rho)\tau_{ij}(t) + \Delta\tau_{ij}(t) \tag{16}$$

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{m} \Delta\tau_{ij}^{k}(t) \tag{17}$$

$\Delta\tau_{ij}$ represented the increment of the pheromone on the path in this circle. In the initial moment, $\Delta\tau_{ij} = 0$. $\Delta\tau_{ij}^{k}(t)$ represented the first ant k release in the circulation process of pheromone on the path the i, j.

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \dfrac{Q}{D_k} & IF\ ant\ k\ pass\ the \\ & path\ i, j\ in\ this\ cycle \\ 0 & o.w \end{cases} \tag{18}$$

Q is a constant, represent the pheromone strength. D_k represents the circle formed by the first ant k times and time units. The ant loop, transfer is determined by the transition probability p_k to that node.

$$p_{ij}^{k} = \begin{cases} \dfrac{\tau_{ij}^{\alpha} \eta_{ij}^{\beta}}{\sum_{s \in allowed_k} \tau_{js}^{\alpha} \eta_{js}^{\beta}} & j \in allowed_k \\ 0 & o.w \end{cases} \quad (19)$$

Among, $allowed_k = \{c - tabu_k\}$ represent the select nodes that ant k set current. $tabu_k$ is taboo table, which records the ant k has been passing nodes and can't walk, used to indicate the memory of the artificial ants. η_{ij} means some heuristic information, in the optimization problem $\eta_{ij} = d_{ij}^{-1}$, α , β embodies the pheromone and the heuristic information for decision-making the impact of ants.

So the probability of XML query into the particle swarm optimization (PSO) and ant colony algorithm, the main process is as follows: first, using particle swarm method, produce a series of random particles, and under the constraint condition, find out the query results of the optimal solution, and records the corresponding path, followed by using ant colony algorithm, solving the pheromone on the path, when carries on the data query again, just the path of the first query has successfully, if not success, again using particle swarm algorithm, generate new particles, and solving the corresponding values, as shown in Figure 2.

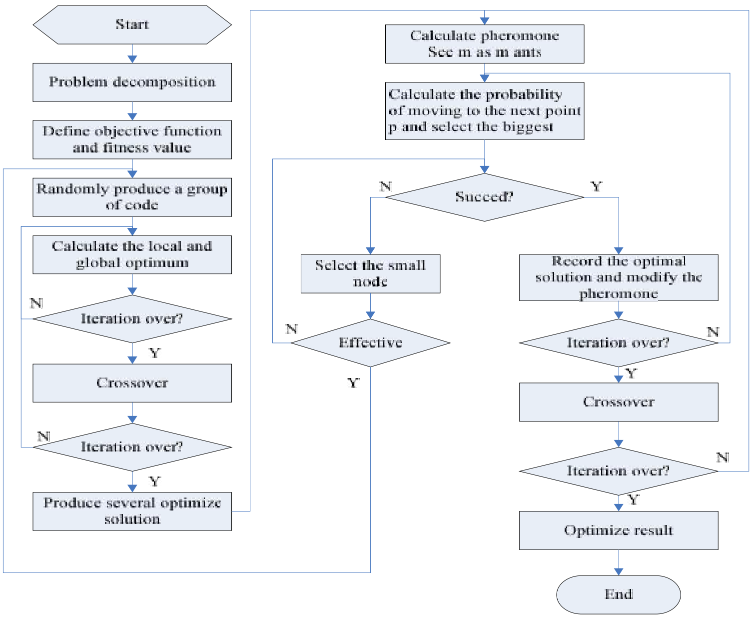


Figure 2. XML Query whole Framework Based on Ant Colony and Particle Swarm Algorithm

4. Simulation and Analysis

4.1. Data Preparation

In order to verify the correctness of the probability query algorithm, and whether can satisfy the condition. The corresponding experiment was carried out, experimental environment under the following conditions: the operating system: professional sp4 version win2000, database: mssql2005, CPU: C - M1.3 GB and 768 MB of memory, hard disk: 80 GB, 5400 RPM, huge amounts of data to survive DataFactory5.2 version. Database table used in this study are for the structure of Table 1, and Table 2. Its record size set is 100, 500, 1000 and 2000. Respectively, four characters of the field length is 2, respectively, to a uniform database for cleaning, query rewriting and probabilistic query rewrite, with P = 1, 2, 3 respectively, key attributes and dynamic properties each of 1.

4.2. Simulation and Analysis

Different combinations of parameters can get different database. Test on a different database will get different results. Here mainly listed several typical schemes of these test results, according to the several groups of test data can be analyzed from a variety of methods for the consistency of database query optimization efficiency.

Table 4. Three Query Relationships Number under the Same Condition

Record number	record number for each group form	Cleaning		rewrite		probability Query		The repeat times of the later two items
		Query	Condition Query	Query	Condition Query	Query	Condition Query	
100	T1=94,T2=93	4200	84	2500	50	801	10	<56
500	T1=341,T2=357	106000	424	19500	78	15916	37	<160*144
1000	T1=509,T2=509	413000	826	38922	78	36058	39	<492*492
2000	T1=623,T2=632	1640000	3288	79880	160	76408	80	<1378*1369

Each group form is record the number.

By Table 4, the proportion of the query frequency of data cleaning write growed much faster than the query rewrite, and the probability of the query accuracy than write faster and more accurate, especially when the precision of the probability, the greater the search range is smaller. The potential for conflict data is less, but because of rewriting query need to duplicate records with probability grouping, so although the accuracy is better than cleaning, but need to query multiple tables of consisting. The price is relatively higher than clean, but more reliability than cleaning, because cleaning will make some useful data loss does not meet the conditions, so the probability of query method can be widely used in intelligent decision making and data mining.

To compare with Reference [4]' dynamic rules with the pruning in the data during the condition of simulation calculate. Query, respectively, three of the same node, according to the proposed ant colony algorithm with hybrid operator is calculated, results as follows:

Table 5. Simulation Results Compared

Times	m=5	m=10	m=50	m=100
Pruning	10	12	65	126
Ant colony	10	12	51	87

5. Conclusion

This paper puts forward an improved algorithm based on probability of XML query, combined with particle swarm optimization and ant colony algorithm. First of all, using particle swarm algorithm and the hybrid method is rapid probability of a set of nodes, and then through the orthogonal operator seeks the optimal solution of ant colony algorithm and global (or local) suboptimal solution path pheromone strength. With the strength of the pheromone size improvement $XPath$ query method, which effectively overcomes the defect of tree query pattern repeated comparison, makes an $XPath$ query is more intelligent and adaptive, which is helpful for the application of XML query and promotion.

Acknowledgment

The authors acknowledge financial support from the Natural Science Foundation of ** City, China (2010A610135), the Innovation and Entrepreneurship Fund of Ningbo City, China (***), the Funds for Creative Research Groups of China (No. 60821063) and 111 Project, China (No. B08015).

References

[1] Barbara D, Garcla-Molina H, Porter D. The Management of Probabilistic Data. *IEEE Trans.Knowl.Data Eng.* 1992; 10: 487-502.

[2] Pittarelli M. An Algebra for Probabilistic Databases. *IEEE Trans. Knowl. Data Eng.* 1994; 6: 293-302.
[3] Yuan Dingrong YanXiaoWei, Chen Hongchao. A new probability data model. *Journal of research in computer application.* 2003: 65-67.
[4] Li Shijun Tan Chengyu, Liu HaiQing etc. A probabilistic relation database system [J]. *Computer engineering*. 2001; 27 (2): 45-46.
[5] M Arenas, L Bertossi, J Chomicki. *Consistent query answers in inconsistent databases*. In Proc. ACM PODS. 1999: 68–79.
[6] P Barcelo, L Bertossi. *Logic programs for querying inconsistent databases*. In Proc. PADL, Springer LNCS 2562. 2003; 208–222．
[7] Surajit Chaudhuri, Gautam Das, Vagelis Hristidis etc. *Probabilistic Ranking of Database Query Results*. Proceedings of the 30th VLDB Conference, Toronto, Canada. 2004; 888-899.
[8] Nilesh Dalvi, Dan Suciu. *Efficient Query Evaluation on Probabilistic Databases*. Proceedings of the 30th VLDB Conference, Toronto, Canada. 2004; 1-12.
[9] Nilesh Dalvi, Dan Suciu. *Answering Queries from Statistics and Probabilistic Views*. Proceedings of the 31st VLDB Conference, Trondheim, Norway. 2005; 805-816.
[10] Maurice van Keulen, Ander de Keijzer, Wouter Alink. A probabilistic XML approach to data integration. *Computer Society*. 2005; 1-12.
[11] Wenzhong Zhao, Alex Dekhtyar, Judy Goldsmith. A framework for management of semistructured probabilistic data. *Journal of Intelligent Information Systems*. 2005; 25(3): 293-332.
[12] Rafael C Carrasco, Juan Ramon Rico-Juan. A similarity between probabilistic tree languages: application to XML document families. *Pattern Recognition*. 2003; 36: 2197-2199.
[13] Cannataro M, Pugliese A. *An XML-Based Architecture for Adaptive Web Hypermedia Systems using a Probabilistic User Model.* Proceedings of the IEEE IDEAS Conference. 2000; 257-265.
[14] Wu Y, Patel JM, Jagadish HV. Structural Join Order selection for XML Query Optimization. Los Alamitos: *IEEE Computer Society.* 2003; 443-454.
[15] Erge Abite boul, Pierre Senellart. Querying and Updating Probabilistic Information in XML. Y.Ioannidis et al.(Eds.). 2006; 3896: 1059-1068.
[16] Hong Zhang, Heng Li, CM Tam. Particle swarm optimization for resource-constrained project scheduling. *International Journal of Project Management*. 2006; 24: 83-92.
[17] Dorigo M, Gambardella LM. Ant Colony System: a Cooperative Learning Approach to the Traveling Salesman Problem. *IEEE Transactions on Evolutionary Computing.* 1997; 1(1): 53-56.
[18] Ding Jianli, Chen Zengqiang, Yuan Zhuzhi. The combination of genetic algorithm and ant algorithm. *computer research and development.* 2003; 40(9): 1351-1356.
[19] TAVAKKOLI-MOGHADDAM R, RAHIMI-VAHED AR. Multi-criteria sequencing problem for a mixed-model assembly line in a JIT production system. Applied Mathematics and Computation. 2006; 181(2): 1471-1481.
[20] FAHIMI-VAHED AR, MIRGHORHANI SM, RABBANI M. A new particle swarm algorithm for a multi-objective mixed-model assembly line sequencing problem. Soft Computing: *A Fusion of Foundations, Methodologies and Applications*. 2007; 11(10): 997-1012.
[21] DONG Qiao-ying, KAN Shu-lin, GUI Yuan-kun, et al. Mixed Model Assembly Line Multi-objective Sequencing Based on Modified Discrete Particle Swarm Optimization Algorithm. *Journal of System Simulation*. 2009; 21(22): 7103-7108.
[22] Eberhart R, Kennedy J. A new optimizer using particle swarm theory. In: Proc of the 6th International Symposium on Micro Machine and Human Science. Piscataway, NJ: IEEE Service Center, 1995: 139~143.
[23] ZHANG Li-Biao, ZHOU Chun-Guang, MA Ming, et al. Solutions of Multi-Objective Optimization Problems Based on Particle Swarm Optimization. *Journal of Computer Research and Development*. 2004; 41(7): 1287-1292.