DOI:10.13718/j.cnki.xsxb.2021.05.025

基于多目标蚁狮优化的众包系统查询优化方案①

刘永涛[1]，　刘　辉[2]

1. 漯河医学高等专科学校 现代教育技术中心，河南 漯河 462001；
2. 河南大学 计算机与信息工程学院，河南 开封 475001

摘要：针对众包系统中的查询优化问题，提出一种基于多目标蚁狮优化的众包系统查询优化方案，用于解决关系数据库系统中查询成本高和延迟时间长的问题. 该方案以延迟和查询成本作为目标函数，在给定用户定义预算约束的情况下，通过蚁狮与轮盘赌策略对蚂蚁种群更新迭代，降序全排蚂蚁和蚁狮的适应度，更新蚁狮种群，采用最优蚁狮更新精英蚁狮，将最后适应值的精英蚁狮作为最优解，从而优化众包系统查询计划，得到低延迟的查询计划，平衡用户的成本和时间要求. 实验结果表明：本文提出的多目标优化方法在众包查询优化中具有一定的优越性，相较于其他算法，该方法在成本和延迟方面都有明显的改善.

关　键　词：众包系统；蚁狮优化算法；查询优化；多目标优化

中图分类号：TP393　　**文献标志码**：A　　**文章编号**：1000－5471(2021)05－0164－06

在数据库管理领域，查询优化(query optimization，QO)是备受关注的热点问题. 查询优化程序通过考虑可能的查询计划来调节最有效的方式用于评估给定的查询计划[1]. 近年来，来自不同资源的大量数据泛滥成灾，使得查询优化成为研究人员的一项艰巨任务. 由于数据库结构复杂，特别是对于不太简单的查询，可以通过不同的方式、数据结构和顺序从数据库中收集查询所需的数据，不同的查询方式通常需要不同的处理时间[2-3].

在众包数据库系统中，用户需要提交基于结构化查询语言(SQL)的查询，然后系统负责查询编译，并生成执行计划和评估众包市场[4-5]. 在这个过程中，由于各种查询计划差异巨大，使得查询优化必不可少. 为了减少查询处理时间和成本，研究人员开发了许多众包查询优化算法. Amini 等[6]提出了一种基于众包的 XML 关键词搜索方法，利用了用户的参与来改进数据库和查询处理的流程. Cincy 等[7]提出一种基于人工蜂群和蝙蝠算法(A-BAT)的众包系统查询优化方法，通过使用 A-BAT 算法计算用户最大访问量来找出优化产品，最大程度地降低成本和等待时间. Rekatsinas 等[8]设计了一个自适应查询框架，通过构造目标域结构和使用排除列表来限制众包查询的重复提取. Bhaskar 等[9]利用鲸鱼优化算法来优化基于位置的查询，并利用众包并行和串行处理来估计查询答案的准确性，降低计算时间和通信成本.

查询优化器通过估计货币成本和延迟时间来确定最佳查询计划. 本文的创新之处在于将查询优化问题视为一个多目标优化问题，采用基于蚁狮优化的多目标优化方法，通过控制质量、最小化成本和时间的方式在众包查询中搜索最佳查询计划. 蚁狮优化算法是一种元启发式算法，蚁狮优化器遵循蚂蚁随机游走、蚁狮建立陷阱、蚂蚁落入陷阱、蚁狮捕捉猎物和重建陷阱 5 个步骤，在控制质量的前提下，以延迟和查询

① 收稿日期：2020－06－10
基金项目：2017 河南省高等教育教学改革研究与实践项目(2017SJGLX159).
作者简介：刘永涛，硕士，讲师，主要从事大数据技术研究.

成本作为目标函数，寻找满足要求的最佳查询计划.

1　众包系统查询优化模型

众包技术作为一种新的计算范式，利用人类智能来解决计算机无法有效完成的问题，用于弥合基于机器计算和基于人类计算之间的鸿沟. 在众包系统中，对于给定的请求者查询，系统首先将查询解析为具有人群驱动操作的查询计划，然后生成要在众包平台上发布的任务，最后收集人群的输入生成结果.

该系统包括用户查询、查询优化、众包执行和众包平台 4 个模块. 首先，用户提交基于 SQL 的查询内容，然后智能查询优化器通过解析后找到合适的查询计划，并最终从查询成本和延迟的角度对其进行正确评估. 查询优化器生成最佳查询计划后，在众包执行模块中执行，生成人类智能任务(human-intelligent tasks，HIT)，并将这些 HIT 转移到众包平台上. 基于从人群中收集的 HIT 答案，执行程序执行查询，并将生成的结果返回给用户.

众包系统中的优化机制大致可以分为基于规则和基于成本两类方法. 基于规则的优化器仅应用一组规则，而不用估计成本来确定最佳查询计划. 尽管基于规则的优化易于实现，但其优化能力有限，且容易出现执行计划无效的问题. 基于成本的优化方法，可以通过估算备选查询计划的成本来评估查询结果，并选择估算成本最低的查询计划. 当前查询优化方法大多采用基于成本的方式进行优化.

2　基于多目标蚁狮优化的 QO 方法

在关系数据库系统中，查询优化在提供声明性查询接口的众包系统中是必需的. 本文考虑了众包系统中几个常用的操作符：CFILL(请求大众填写数据库中缺少的值)；CSELECT(要求大众筛选满足某些约束的项)；CJOIN(根据某些条件利用大众匹配项)；CSORT(根据条件对选择进行排序). 针对 SQL 语言的局限性，采用一种多目标蚁狮优化的查询优化方法，该方法以延迟和查询成本作为目标函数，在给定用户定义的预算约束情况下，找到低延迟的查询计划，从而很好地平衡用户的成本和时间要求. 其中，优化查询涉及的成本为所有操作符的总体成本.

本文考虑两个优化目标，第一个仅考虑货币成本：给定一个查询 Q，目的是找到一个使货币成本最小的查询计划 P_Q^*

$$P_Q^* = \arg_{PQ}\min cost(P_Q) \tag{1}$$

由于目标 1 下的最优计划可能需要很长时间才能完成，因此需要引入第二个目标，该目标考虑了延迟，目的是在用户提供成本预算下找到一个低延迟的查询计划：给定一个查询 Q 和成本预算 C，它找到一个查询计划 P_Q^*

$$\begin{aligned} &P_Q^* = \arg_{PQ}\min latency(P_Q) \\ &s.t.\ C(P_Q^*) \leqslant C \end{aligned} \tag{2}$$

式(1)、式(2)中：$cost$ 和 $latency$ 分别表示查询计划 P_Q 的查询成本和延迟时间. 如果有多个延迟最小的计划，它将找到成本最低的计划.

本文算法提供了一个优化框架. 该算法以众包查询 Q 和成本预算 C 作为输入，并生成优化的查询计划 P^*. 算法考虑以下两种情况：如果未指定成本预算，则以成本最小化为目标搜索 Q 的查询计划. 否则，算法将以成本限制的延迟最小化为目标函数. 优化框架分为 4 个主要步骤：收集数据生成数据库，生成用户查询、优化用户查询以及将其应用于众包系统.

2.1　数据收集

在基于查询的众包系统初始阶段，本文采用输入数据为加州大学欧文分校(UCI)汽车数据集内的数据[10]. 对于 UCI 数据集，首先使用数据集中的 205 辆汽车参数来生成车辆之间的关系；其次将 VEHICLE 中的每个元组复制 20 次来生成关系 IMAGE，并添加颜色和质量这两个属性，其值是随机生成的. 同理，

将VEHICLE中的每个元组重复10次生成关系REVIEW，并添加具有随机生成的值的属性情感. 在数据集UCI上实现一个模拟众包环境，此环境具有UCI的完整数据库知识. 当点击到达时，它会搜索完整的数据库并将正确的答案返回给众包执行者.

2.2 查询优化

数据收集后，将收集的数据以SQL格式存储. 由于存储的SQL数据结构复杂性不断增长，使得检索所需数据的时间也随之增加. 目前，手工编写更复杂文档的查询是一项非常容易出错且繁琐的任务. 这些问题尤其在数据仓库中被引起关注，在数据库中文档从不同的来源收集，其结构略有不同. 因此，需要一种有原则的方法来协助用户建立查询.

在所提出的方法中，使用关系数据模型. 首先，数据被指定为一组关系$R=\{R_1, R_2, \cdots, R_{|n|}\}$，这些关系可以由操作人员指定，并且能够被众包用户查询. 每个关系R_i包含一组描述元组性质的属性$\{A_1^i, A_2^i, \cdots, A_m^i\}$. 与传统数据库不同，在执行众包之前元组的某些属性是未知的. 在查询计划中，应用SELECT,JOIN和FILL这3种类型的运算符，然后使用蚁狮优化器执行查询优化，以便为查询检索提供更好的解决方案. 蚁狮优化算法是一种基于种群的算法，它有助于确定随机解集的最优解. 当用户遇到困难时，查询优化器能够给出问题的解决方案. 这样可以提高效率，并为问题提供更好的解决方案.

2.3 蚁狮优化算法

蚁狮优化算法(ant lion optimization，ALO)是一种群智能优化算法[11]，该算法模拟了自然界中蚁狮捕食蚂蚁的行为活动，ALO算法具有调节参数少、寻优精度高、不易陷入局部最优等优点. 蚁狮优化算法中具有蚂蚁和蚁狮两个群体，寻优过程可分为蚂蚁随机游走、蚁狮建立陷阱、蚂蚁落入陷阱、蚁狮捕捉猎物和重建陷阱5个步骤. 蚂蚁随机游走时的公式可以定义为

$$X(t)=[0, cs(2r(t_1)-1), cs(2r(t_2)-1), \cdots, cs(2r(t_n)-1)] \tag{3}$$

式(3)中：n表示最大迭代次数，t表示随机游走的步骤，cs表示计算累积和. $r(t)$表示随机函数，其定义为

$$r(t)=\begin{cases}1 & rand>0.5\\ 0 & rand\leqslant 0.5\end{cases} \tag{4}$$

式(4)中：$rand$表示在(0，1)内均有分布的随机数. 为了保证蚂蚁在可行域范围内行走，将蚂蚁的随机游走行为归一化处理

$$X_i^t=\frac{(X_i^t-a_i)\times(d_i^t-c_i^t)}{b_i-a_i}+c_i^t \tag{5}$$

式(5)中：c_i^t和d_i^t分别表示蚂蚁第i分量在第t次迭代时的最小值和最大值，a_i和b_i分别为蚂蚁第i分量的最小值和最大值. 同时，蚂蚁随机游走也受到蚁狮陷阱的影响，是在选定的蚁狮周围运动.

$$\begin{cases}c_i^t=Antlion_j^t+c^t\\ d_i^t=Antlion_j^t+d^t\end{cases} \tag{6}$$

式(6)中：$Antlion_j^t$表示第j个蚁狮在第t次迭代时的位置，c^t和d^t分别为第t次迭代时的最小值和最大值.

ALO算法为了提高收敛速度，寻求得到最优解，将蚂蚁随机游走的范围随迭代次数增加而逐渐降低.

$$\begin{cases}c^t=\dfrac{c^t}{I}\\ d^t=\dfrac{d^t}{I}\end{cases} \tag{7}$$

式(7)中：$I=1+10^{\omega}\times\dfrac{t}{n}$，$\omega$表示随迭代次数而动态调整的参数.

蚁狮捕食运动到陷阱底部的蚂蚁后，根据式(2)中多目标适应度值更新位置为

$$Antlion_j^t=Ant_i^t, \text{若 } f(Ant_i^t)<f(Antlion_j^t) \tag{8}$$

式(8)中：Ant_i^t表示第i个蚂蚁在第t次迭代时的位置. 蚁狮更新完位置后，重新构造陷阱，用于捕获下一只

蚂蚁. 选取并保存每次迭代时的最佳蚁狮作为精英蚁狮，精英蚁狮在ALO算法迭代过程中可以影响所有蚂蚁的活动. 为了避免算法陷入局部最优，ALO 算法通过轮盘赌选择和随机游走确定蚂蚁的位置，即

$$Ant_i^t = \frac{R_A^t + R_E^t}{2} \tag{9}$$

式(9) 中：R_A^t 表示由轮盘赌在第 t 次迭代选择的蚁狮周围随机游走产生的值，R_E^t 表示在第 t 次迭代精英蚁狮周围随机游走产生的值.

3　实验与结果分析

为了验证算法的性能，所有实验在 MATLAB 2018a 中进行，运行环境配置为 Net Beans 6.2 版本，3.0 GHz 英特尔 i5 处理器，2 TB 硬盘和 16 GB RAM 工作站. 为了评估本文算法的效率和有效性，从精度、货币成本和延迟等方面考察本文方法的性能，并将实验测试结果与 A-BAT 算法[7]、WOA 算法[9]和 CDB 算法[12]等查询优化算法进行比较.

为了验证选择、联接、复杂查询和排序等几个常用操作符在众包查询优化算法中的有效性，在实验中使用线性价格函数 $f=b+wx$，其中 b 和 w 分别表示基本费用和增量费用. 对于 CSELECT，CJOIN 和 CSORT，将 b 和 w 都设置为 \$ 0.005，而对于 CFILL，将 b 和 w 分别设置为 \$ 0.01 和 \$ 0.002.

3.1　性能指标

绩效衡量被定义为对结果的衡量，它提供了一个关于系统有效性和效率的可靠信息. 系统输入值和输出值之间的关系可以通过使用精度、货币成本和延迟等性能指标来评估.

计算众包系统中查询优化货币成本的公式用数学表示为

$$Cost(P_Q) = \sum\nolimits_{o \in O} cost(o) \tag{10}$$

式(10) 中：$cost(o)$ 表示执行每个运算符的成本.

由于众包需要时间，因此自然会引入延迟来量化查询评估的速度，本文将查询计划 P_Q 的等待时间 $L(P_Q)$ 衡量为 P_Q 众包执行中使用的迭代次数. 同时，如果存在垃圾邮件发送者或恶意工作人员，众包可能会产生相对较低质量的结果，甚至产生噪音. 因此，精度被视为衡量众包结果质量的另一个重要性能指标.

3.2　实验分析

在实验分析中，利用 UCI 汽车数据库对本文算法与现有算法进行了性能评价比较. 表 1 和表 2 给出了不使用预算和使用预算时本文算法与其他算法的测试结果.

表 1　不使用预算的成本对比

选择条件	A-BAT	WOA	CDB	本文算法
2	25.1	33.6	37.2	22.8
3	30.3	39.1	42.9	25
4	50.8	55.6	61.7	44.2
5	55.9	64.4	70.3	46.7
6	66.1	75.3	87.6	61.8

从表 1 中可以看出，选择条件的数量为 2～6 个，并且为每个选择条件随机创建 10 个查询并计算平均成本. 在这种情况下，本文方法性能最高，在所有设置中产生了最低的成本. 在表 2 中，对本文方法进行了预算限制评估，预算限制在 80～180 \$ 范围内，改变众包的预算是为了要求大众在预算限制内找到尽可能多的答案，在这里查询被考虑为 6 个选择条件，并与 A-BAT、WOA 和 CDB 的查询优化器进行比较. 首先按选择顺序的升序排列选择条件，然后将其随机打包分组. 从表 2 中可以看出，本文提出的多目标优化算法具有更低的代价. 同时，图 1 给出了不同预设限制下的精度对比，从图 1 中可以看出，在每种预算下，所有优化方法的精度都在 95%以上，都能够达到优化的目的，但是通过对比发现，本文算法的精度最高，几乎达到 99%的精度，优于其他方法.

表2 使用预算的成本对比

用户成本/$	A-BAT	WOA	CDB	本文算法
80	65	69	72	56
100	68	80	83	61
120	74	82	90	69
140	77	79	100	71
160	77	80	103	70
180	94	105	122	89

图2给出了在预算限制条件下不同算法的延迟结果. 在这个实验中，考虑了6种选择条件下的查询，从图2中可以看出，本文所提出的多目标优化算法显著减少了延迟，这是由于本文所提出的技术优势产生了最佳分组，它可以明智地确定哪些条件可以分组以减少延迟，并以最佳顺序应用这些条件.

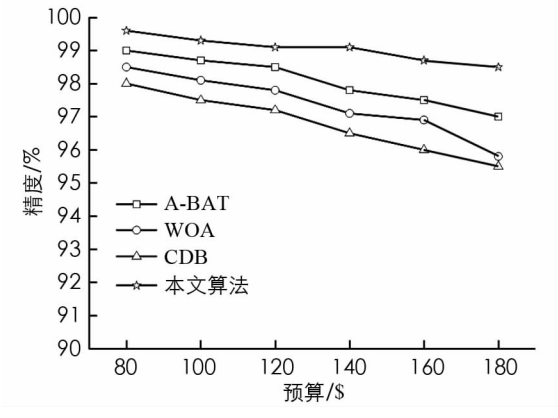

图1 不同预算下的精度对比

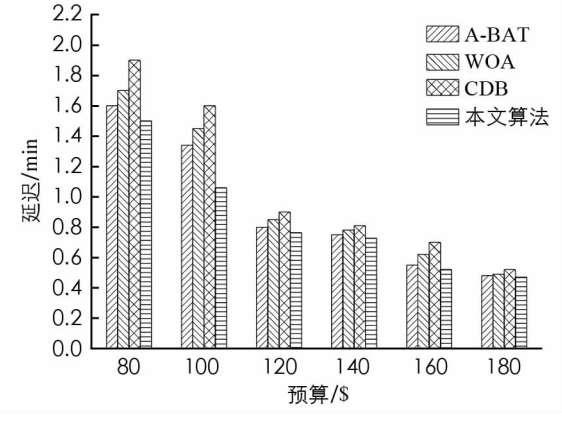

图2 不同预算下的延迟对比

4 结 语

本文针对众包系统查询优化中存在的查询成本高和延迟时间长的问题，提出一种基于多目标蚁狮优化器的查询优化方案，该方案以最小化延迟和查询成本作为目标函数，通过优化器找到满足固定预算条件下的低延迟查询计划，如果存在多个延迟最小的计划，则选择成本最低的计划作为最佳查询计划. 本文所提出的方案考虑了成本和延迟直接的平衡并支持选择、联接和复杂查询等多个众包操作符. 实验结果表明，本文所提出的众包查询优化方案比其他算法更有优势，明显降低了查询优化的成本.

参考文献：

[1] PANAHI V, NAVIMIPOUR N J. Join Query Optimization in the Distributed Database System Using an Artificial Bee Colony Algorithm and Genetic Operators [J]. Concurrency and Computation: Practice and Experience, 2019, 31(17): e5218.

[2] 邹承明，谢 义，吴 佩. 基于Greenplum数据库的查询优化 [J]. 计算机应用，2018，38(2)：478-482.

[3] PANG Z F, WU S, HUANG H C, et al. AQUA+: Query Optimization for Hybrid Database-MapReduce System [C]// 2019 IEEE International Conference on Big Knowledge (ICBK). Beijing: IEEE, 2019.

[4] ZHANG D T, WEN S T, CHEN F, et al. Spatial Crowdsourcing Based on Web Mapping Services [J]. World Wide Web, 2020, 23(1): 631-648.

[5] CHAI C L, FAN J, LI G L, et al. Crowdsourcing Database Systems: Overview and Challenges [C]//2019 IEEE 35th International Conference on Data Engineering (ICDE). Macao: IEEE, 2019.

[6] AMINI L M, KEYVANPOUR M. A Crowdsourcing-Based Approach for Efficient XML Keyword Search [C]//2019 5th International Conference on Web Research (ICWR). Tehran: IEEE, 2019.

[7] CINCY W C, JEBA J R. A Method of A-BAT Algorithm Based Query Optimization for Crowd Sourcing System [J].

International Journal of Intelligent Systems and Applications，2018，10(3)：33-40.

[8] REKATSINAS T，DESHPANDE A，PARAMESWARAN A. CRUX：Adaptive Querying for Efficient Crowdsourced Data Extraction [C]//CIKM '19：Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing：ACM，2019.

[9] BHASKAR N，KUMAR P M. Optimal Processing of Nearest-Neighbor User Queries in Crowdsourcing Based on the Whale Optimization Algorithm [J]. Soft Computing，2020，24(17)：13037-13050.

[10] WANG Y B，XU W. Leveraging Deep Learning with LDA-Based Text Analytics to Detect Automobile Insurance Fraud [J]. Decision Support Systems，2018，105：87-95.

[11] ASSIRI A S，HUSSIEN A G，AMIN M. Ant Lion Optimization：Variants，Hybrids，and Applications [J]. IEEE Access，2020，8：77746-77764.

[12] LI G L，CHAI C J，FAN J，et al. CDB：a Crowd-Powered Database System [C]. Rio de Janeiro：2018 44th International Conference on Very Large Data Bases(VLDB)，2018.

Query Optimization Scheme in Crowdsourcing System Based on Multi-Objective Ant-Lion Optimization

LIU Yong-tao[1]，　LIU　Hui[2]

1. *Modern Education Technology Center，Luohe Medical College，Luohe Henan 462001，China*；

2. *School of Computer and Information Engineering，Henan University，Kaifeng Henan 475001，China*

Abstract：Aiming at the problem of query optimization in crowdsourcing system，a query optimization solution for crowdsourcing system based on multi-objective ant lion optimization has been proposed to solve the problem of high query cost and long delay time in relational database system. The scheme takes delay and query cost as the objective functions. Given the user-defined budget constraints，the ant population is updated iteratively through the ant lion and roulette strategy，and the ant and the lion's fitness are ranked in descending order. The lion population adopts the optimal ant lion to update the elite ant lion，and uses the elite ant lion with the last fitness value as the optimal solution，thereby optimizing the query plan of the crowdsourcing system，obtaining a low-latency query plan，and balancing the cost and time requirements of users. The experimental results show that the proposed multi-objective optimization method has some advantages in crowdsourcing query optimization. Compared with other algorithms，this method has obvious improvement in cost and delay.

Key words：crowdsourcing system；ant-lion optimization algorithm；query optimization；multi-objective optimization

责任编辑　夏　娟