文章编号:1006-9348(2021)08-0381-05

基于群体智能的半结构化数据查询优化算法

高俊杰[1],杨 帆[2]

(1.山西大学计算机与信息技术学院,山西太原 030006;
2.山西医科大学计算机教学部,山西太原 030001)

摘要:传统算法不能有效结合半结构化数据特征,在进行算法运行过程中所查询的数据量较少,且时间较长。于是基于群体智能研究了一种新的半结构化数据查询算法。采用粒子群优化算法建立半结构化数据查询模型,标记空间内的数据特征,运用中心-离散算法对计算模型数据中的不同粒子类型进行查询,在不同范围内实现对数据的查询与搜索。采用映射方法形成数据查询模型集合,并利用映射关系与有向图中的数据内容建设半结构化数据模型储存空间,利用标签树区分数据结构与数据层次,实现对半结构化数据的查询优化。实验结果表明,所提算法的数据查询量更多,查询时间更短,说明上述方法优化能力更强。

关键词:群体智能;数据查询;粒子群体优化算法;半结构化模型;标签树

中图分类号:TP301 **文献标识码**:B

Semi-Structured Data Query Optimization Algorithm Based on Swarm Intelligence

GAO Jun-jie[1], YANG Fan[2]

(1. School of Computer & Information Technology, Shanxi University, TaiyuanShanxi030006, China;
2. Department of Computer Teaching, Shanxi Medical University, TaiyuanShanxi030001, China)

ABSTRACT: Traditional algorithms can not effectively combine the characteristics of semi-structured data, so the amount of data queried in the process of algorithm running is less and the time is longer. In this regard, we studied a new semi-structured data query algorithm based on swarm intelligence. First of all, particle swarm optimization algorithm was used to found semi-structured data query model and label the data features in the space. Secondly, the center discrete algorithm was applied to query different particle types in the computational model data, realizing the query and search of data in different ranges. Then, based on the mapping method, the data query model set was established, and the semi-structured data model storage space was constructed by using the mapping relationship and the data content in the directed graph. Eventually, the tag tree was used to distinguish the data structure and data hierarchy to realize the query optimization of semi-structured data. The experimental results show that the algorithm has more data queries and shorter query time, indicating that the optimization ability of the method is stronger.

KEYWORDS: swarm intelligence; data query; particle swarm optimization algorithm; semi-structured model; tag tree

1 引言

随着计算机技术的不断发展,人类已经逐渐走向全面数字化时代,人们在各个领域中都开始应用数字化技术,为此需要研究能够高效率处理大数据的方法,解决数据智能化等问题[1-3]。为了能够区分不同类型的数据,将网络数据按照信息结构进行划分,为计算机系统提供灵活的数据转换平台[4]。半结构化数据由于其组成信息结构的不完整性,需要不断地对数据结构进行修改,再根据半结构化数据的特征完成数据查询、数据采集、数据分析等内容,半结构化数据查询是较为复杂的一项基础工作,只有从多角度完成数据分析并及时发现潜在数据才能稳定支持计算机系统运行[5,6]。

基金项目:国家自然科学基金重点项目(61936012):基于语言认知机理的汉语框架语义计算研究
收稿日期:2020-09-11

文献[7]提出一种基于路径索引的密集邻域图数据查询方法,该方法通过分析顶点密集领域数据特点,设计顶点密集邻域的数据查询模式,运用 B + 树方法对数据查询路径进行索引与存储。实验结果表明,该方法能够提升数据查询与处理性能,但是存在数据查询量较少的问题。文献[8]提出了基于 Python 的协议栈软件内部数据查询方法,该方法通过 Linux 系统与 Windows 系统的兼容,实现直接使用命令在屏幕上输出所查数据结构中的所有数据。实验结果表明,该方法数据查询精度较高,但是耗时较长。除此之外,文献[9]还提出了基于符号语义的不完整数据聚集查询处理算法,该算法对传统关系数据库模型进行了扩展,实现对不完整数据聚集的区间估计,并在该区间内实现对数据的查询。实验结果表明,该算法数据查询速度较快,但是得到的数据量较少。

针对现有方法存在的问题,提出基于群体智能的半结构化数据查询优化算法。

2 基于粒子群优化算法的半结构化数据识别

采用粒子群优化算法作为半结构化数据查询模型的优化算法,应用该算法的环境移动特征能够更加精准、快速地在一定范围内寻找可靠数据,可以在三维空间中建立数据质量与数据规模的位置检索空间,在空间内建立多种数据结构代表粒子,粒子在半结构化模型中的飞行过程称为查询过程,具体的飞行轨迹与飞行终点是根据粒子数据的自身经验来完成动态调整的,具有不确定性[10]。三维空间粒子分布如图 1 所示。

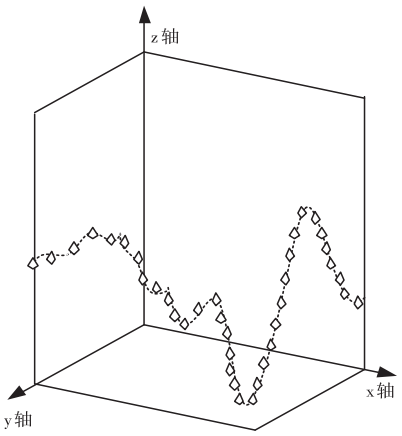

图 1　三维空间粒子分布

根据图 1 可知,基本的粒子群算法应用需要建设一个三维立体空间为粒子群提供飞行场地,每个粒子所在的粒子种群可以代表在三维空间中的精准位置,粒子飞行的速度与种群迭代速度主要通过设定位置参数的方式进行表示,基本粒子群中的粒子超出标准飞行速度,则三维立体空间内的维度参数会根据超出的飞行速度重新定义,制定超出空间边界的参数值[11]。标准的粒子群算法基于基本粒子群算法的三维立体空间,可以在粒子飞行过程中完成算法的惯性参数计算,相对于基本粒子群算法有着较强的空间查询能力,此算法在半结构化查询模型中主要通过线性递增的方式调节数据的检索范围,实现了全方面的数据结构探索。在结构模型中进行深度的数据种子探索,为粒子群的收敛速度提供保障,粒子飞行的基本惯性参数可以在线性关系的基础上引用常数因子,粒子飞行速度与常数因子呈现权重关系。离散粒子群算法是粒子群优化算法中的高阶算法,能够对持续性函数提出优化解法,粒子在每个空间维度中的飞行状态都需要经过向量标准进行评判,若粒子的飞行速度没有受到半结构化数据查询模型的限制,粒子群则建立随机数据检索范围,针对更深层次的算法搜索空间建立粒子飞行节点[12]。离散粒子分布状态如下图 2 所示。

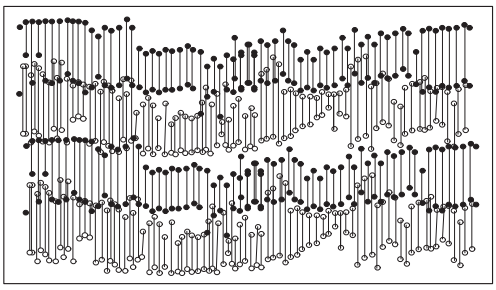

图 2　离散粒子分布状态

中心学习算法建立在粒子群中的共享中央位置,此算法能够为整个粒子群中的每个粒子解决粒子飞行层面问题,当粒子种群中的部分个体产生位置聚拢情况时,中心学习算法能够协同粒子参数与种群的数据查询机制,为该种群的局部空间设定阈值。在中心学习算法中的迭代关系主要依靠粒子的维度空间运转速度,中心学习算法的迭代运行公式为

$$\begin{cases} v(t+1) = \omega v(t) \\ x(t+1) = \omega(t) \end{cases} \tag{1}$$

式中,v 代表普通粒子在维度空间内的飞行速度;x 代表中心学习算法中的精英粒子在维度空间中的飞行速度;t 代表维度空间内的种群飞行时间;ω 代表粒子惯性权重。

粒子离散过程如图 3 所示。

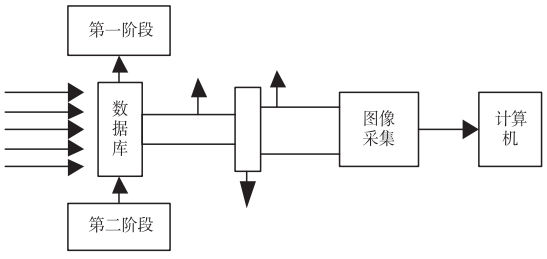

图 3　粒子离散过程

离散学习算法是一种能够在数据查询模型中进行大范

围数据查询的搜索算法,中心学习算法主要用于计算模型数据中的精英粒子,离散学习算法主要用于计算模型数据中的普通粒子,普通粒子需要对应每个维度空间阶层,在离散学习层次上需要具备不同的粒子个体。离散学习算法的迭代运行公式为

$$\begin{cases}\lambda(t+1) = \omega_0(t'+1)\\ \partial(t+1) = \omega_0\end{cases} \tag{2}$$

式中,λ 代表粒子在空间中的飞行速度;t'代表粒子在空间中的总飞行时间;ω_0 代表粒子学习对象的权重值;∂代表粒子的分布离散系数。

为了解决粒子群在半结构化查询模型中的粒子收敛不均匀的现象,在模型中进行维度协同,应用粒子的自身参数设定维度阈值,根据每种类型的粒子参数进行维度内的数据更新。不同维度下的离散数据如图 4 所示。

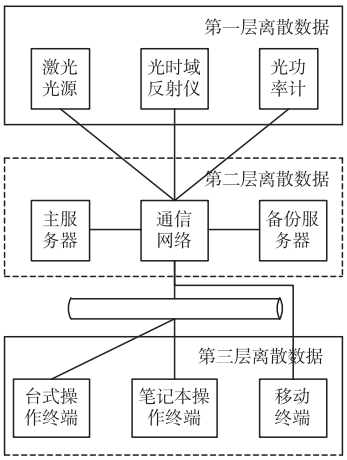

图 4　不同维度下的离散数据

观察图 4 可知,假设一定范围内的结构模型中维度为 N,则粒子种群中每个粒子可以被描述的向量需要进行维度划分,在粒子运行的周期内完成粒子与固定维度的关联。由于维度空间内的位置划分具有一定的随机特性,所以需要每层粒子在飞行前进行随机采样,进行不同维度空间的粒子协同,确保每个粒子在随机维度中都能完成半结构化数据识别。

3　基于群体智能的半结构化数据查询优化算法

以半结构化数据识别结果为基础,进行半结构化数据查询优化。传统半结构化数据模型的设计不考虑数据的查询路径以及数据查询方法等问题,主要设计半结构化数据模型的整体构架以及数据储存空间,而本文在半结构化数据模型建设过程中添加数据导向体系,能够将查询数据在有向图中进行表示,可以在不同数据环境中与其它种类的查询数据相互转换。离散数据采集过程如图 5 所示。

数据标记技术随着数据运用功能的不断丰富,在目前的

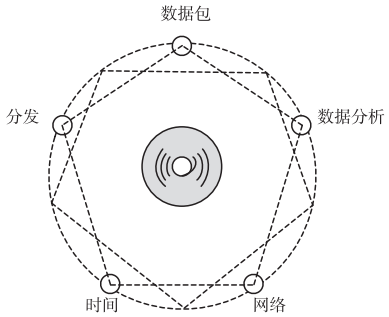

图 5　离散数据采集过程

数据传输领域、数据识别领域、数据采集、查询等领域均有所运用,数据标记技术应用在空间结构的半结构化数据查询模型中,主要体现在数据结构树中,在数据树中首先建立查询对象支路,在每条查询支路中安装编程接口,再通过文档接口映射形成一套数据查询模型集合,集合的运行是经过应用程序对查询文档的数据属性提取后实现的。

随着半结构化数据模型中的数据表现形式越来越丰富,数据查询模型中的数据文档对数据挖掘需求越来越大。需要对数据进行深入挖掘,深入挖掘示意图如图 6 所示。

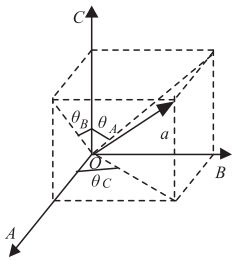

图 6　数据深入挖掘示意图

图 6 中,α、θ 分别表示不同区域的数据挖掘方向。被标记的半结构化数据模型还具有灵活性强的特点,能够在任意时间内完成数据查询程序的更改,数据查询结构内部也在不断地进行数据库更新工作,还能够在同一查询阶段建立多种版本的数据文档,为动态查询程序增加可利用点。由于数据自身具备流动特性,所以本文针对频繁变化的数据内容建立标记特征的半结构化模型,在已知的动态数据结构基础上,开发文档版本与文档版本之间的了半结构化模型。设计子半结构化模型的文档版本查询阈值,不断累积查询用户与查询参数之间的融合信息。优化时间如图 7 所示。

根据图 7 可知,T 代表优化时间,S_1 代表优化距离。半结构化数据查询模型中常常出现模糊数据的拓展问题,通过统计不同数据库中的一些不确定性信息可知,这些模糊数据在模型中容易与其它类型查询数据结合成为随机数据集合或粗糙数据集合,不方便查询运行程序的管理。为此,基于群体智能标签树研究不确定性查询数据的节点关系,建立基

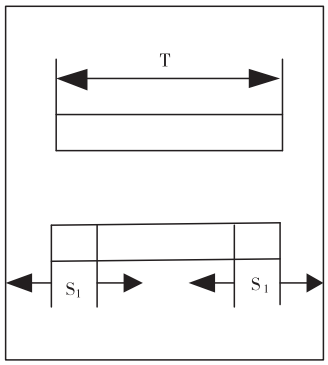

图7　优化时间

本模型,优化数据查询过程中的模糊数据与其它数据建立不确定关系问题。基于群体智能的半结构化数据查询优化过程如图8所示。

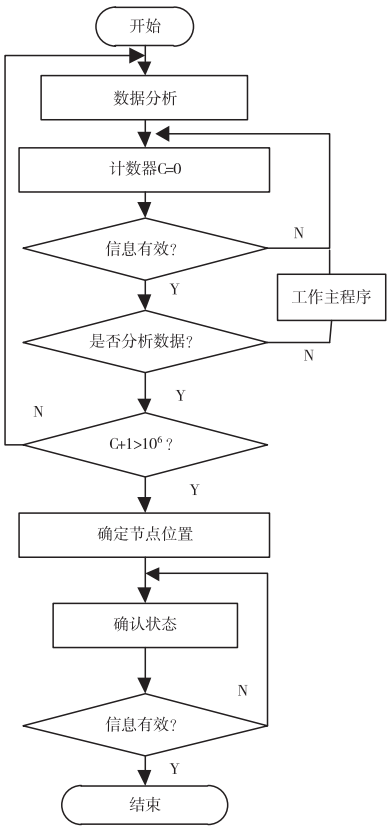

图8　半结构化数据查询优化流程

标签树在半结构化数据查询模型中关联着每层数据结构中的节点,区分数据结构与数据层次,查询标签树的内部数据信息时可以采用数学归纳法建立数据模型与标签数据的等价关系,要求每层次的数据结构与节点中的数据可进行区分。设定标签树在数据模型中的应用语言可以跨平台使用,在传统数据系统 d 基础上增加语言的分析能力,简约化分析标签树语言的精准性。定义标签树信息在数据结构中的决策规则为:

$$\begin{aligned} DR = &(l \rightarrow ac) \wedge (l_0 \rightarrow a_0c_0) \\ &\wedge \cdots \wedge (l_n \rightarrow a_nc_n) \Rightarrow \\ &(l \rightarrow ad) \wedge (l_0 \rightarrow a_0d_0) \\ &\wedge \cdots \wedge (l_n \rightarrow a_nd_n) \end{aligned} \tag{3}$$

式中,$(l \rightarrow ac)$代表标签树数据节点中的决策点;$(l \rightarrow ad)$代表模型决策后的数据关联内容,主要用于半结构化数据决策与分化。

依据式(3)所示的决策规则,确定数据关联关系,从而实现半结构化数据的查询优化。

4　实验测试

为了评估本文研究的基于群体智能的半结构化数据查询优化算法的有效性,设定实验测试环节。选用6个基本函数作为粒子群优化算法的综合性能基准测试函数,前两个函数每个函数中只有一个峰值,后四个函数具有多个峰值,多峰值函数用来检测算法在半结构化模型中的数据查询量,单峰函数用来检测算法在半结构化数据模型中的查询功能强度。

实验中首先测定中心-离散学习算法,将算法代入函数中重复查询数据20次,记录每次查询结果,计算查询结果的均值、方差、极差等值。应用中心-离散学习算法对多峰函数的数据收敛效果进行确认,寻找算法中的函数最佳峰值,并在多峰值的环境中寻找周期性的函数最优解。表1为各函数在算法中的参数设定值:

表1　参数设定值

测试函数	参数值
CDPOS(F1)	$\omega=0.31,\gamma=0.40$
LPSO(F2)	$\omega=0.25,\gamma=0.27$
SPSO(F3)	$\omega=1.05,\gamma=4.36$
FIPSO(F4)	$\omega=0.80,\gamma=0.40$
FIPS(F5)	$\omega=0.52,\gamma=1.02$
FDRPSO(F6)	$\omega=0.90,\gamma=0.35$

实验中,为了保障不同函数在算法中的同步性,需要在粒子种群中设置相同维度的迭代运算,每次完成迭代运算的函数会统计自身的目标函数值与峰值在函数环境中的最佳位置。实验中还应用维尔克森双尾秩和检验方法对实验结果进行验证,图9为本文算法对不同测试函数的数据查询结果。

根据图9可知,本文选用的粒子群优化算法不仅可以实现对半结构化数据的查询,而且随着迭代次数的增长,数据查询量整体上呈现出逐渐增长的趋势,提高了维度空间内非结构化数据的查询深度。

为了进一步验证本文算法的有效性,对比基于路径索引

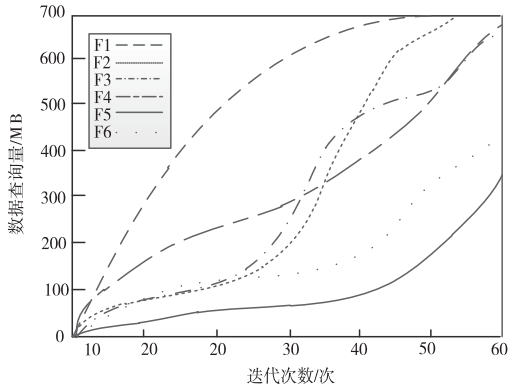

图 9　不同方法数据查询量对比结果

的密集邻域图数据查询方法(文献[7]方法)和基于 Python 的协议栈软件内部数据查询方法(文献[8]方法)与本文算法的数据查询优化时间,结果如下表 2 可知。

表 2　数据查询优化时间对比结果

实验次数/次	查询时间/s		
	文献[7]方法	文献[8]方法	本文算法
1	15.25	17.22	8.71
2	16.85	18.34	6.73
3	19.33	19.44	7.31
4	17.36	18.21	8.22
5	15.24	17.52	8.47
6	14.33	17.36	8.25
7	12.36	18.96	9.37
8	11.02	19.44	8.66
9	18.36	17.02	7.25
10	14.25	16.54	9.42

根据表 2 中的数据可知,本文提出的基于群体智能的半结构化数据查询优化算法优化时间小于传统优化算法。这是由于提出的算法引入了数据分析策略,具有很好的集中性,可以在短时间内完成数据的分析。

综上所述,提出的基于群体智能的半结构化数据查询优化算法优化能力更强,优化效果更好,更适用于实际的数据优化工作。

5　结论

半结构化数据查询模型自身具有结构特性,相对于其它结构的数据查询模型需要不断进行数据修复与结构构建,本文基于群体智能方法,选用粒子群优化算法对传统半结构化数据查询模型进行改良,从数据查询模型的结构、数据查询模型维度空间以及数据查询方式作为切入点,增强半结构化数据查询模型的可靠性与高效性。实验结果表明,该算法具有半结构化数据查询量大和数据查询时间较短的优势。

参考文献:

[1] 程适,王锐,伍国华,等. 群体智能优化算法[J]. 郑州大学学报(工学版), 2018,39(6):1-2.

[2] 薛猛,姜淑娟,王荣存. 基于智能优化算法的测试数据生成综述[J]. 计算机工程与应用, 2018,54(17):16-23.

[3] 徐立鑫,吴化尧. 基于群体智能的软件工程方法综述[J]. 计算机研究与发展, 2020,57(3):487-512.

[4] 秦映波,曹步清,邓春晖. 一种基于竞争型群体优化的数据聚类方法[J]. 计算机与现代化, 2019,281(1):79-83,104.

[5] 罗希意,霍晓阳,傅洛伊. 基于窗口函数和分布式集群的可视化学术搜索系统数据查询优化[J]. 上海交通大学学报, 2019,53(8):92-96.

[6] 饶小康,马瑞,张力,等. 基于人工智能的堤防工程大数据安全管理平台及其实现[J]. 长江科学院院报, 2019,36(10):104-110.

[7] 段慧芳,汤小春. 基于路径索引的密集邻域图数据查询方法研究[J]. 计算机应用研究,2018,35(12):3738-3742,3751.

[8] 蒋玉玲,肖亚楠,方涛. 基于 Python 的协议栈软件内部数据查询方法[J]. 光通信研究, 2018,44(2):29-31.

[9] 张安珍,李建中,高宏. 基于符号语义的不完整数据聚集查询处理算法[J]. 软件学报, 2020,31(2):406-420.

[10] 崔建双,车梦然. 基于多分类支持向量机的优化算法智能推荐系统与实证分析[J]. 计算机工程与科学, 2019,41(1):153-160.

[11] 乔少杰,韩楠,金澈清,等. 基于 Multi-Agent 的分布式文本聚类模型[J]. 计算机学报, 2018,41(8):1709-1721.

[12] 罗玉梅,张德宇极值扰动优化隐私框架的频繁项数据查询研究[J]. 计算机仿真,2020,37(10):403-406.

[作者简介]

高俊杰(1984-),男(汉族),山西浑源人,硕士,实验师,研究方向:智能信息处理、计算机网络。

杨　帆(1987-),男(汉族),山西新绛人,硕士,实验师,研究方向:数据挖掘、形式概念分析、信息智能处理。