doi：10.13413/j.cnki.jdxblxb.2023103

面向用户偏好的动态网页数据交互式查询算法

赵红梅，肖 明，白 宇，王 磊
（黑龙江八一农垦大学 现代教育技术与信息中心，黑龙江 大庆 163316）

摘要：为提高网页数据查询速度、精度及工作效率，提出一种面向用户偏好的动态网页数据交互式查询算法. 首先，构建用户偏好模型，增加偏好组合的演化个体适应性，综合计算适配值；其次，为防止数据冗余和重复，基于兴趣相似性，分离相似度高的查询数据和重复数据，识别出网络数据的性质；最后，利用粒子群优化算法寻找最优的动态网页数据交互式查询方案. 实验结果表明：在数据集基数影响下，该算法的查询结果集质量在 0.95 以上；在查询最大维数影响下，该算法的查询结果集质量在 0.96 以上，表明其查询使用时间短、结果集精度高、自适应能力强.

关键词：用户偏好模型；动态网页数据；数据交互式查询；粒子群优化算法；空间维度
中图分类号：TP311 **文献标志码：**A **文章编号：**1671-5489(2024)02-0417-06

Interactive Query Algorithm for Dynamic Web Page Data Based on User Preference

ZHAO Hongmei, XIAO Ming, BAI Yu, WANG Lei
(*Center of Modern Educational Technology and Information, Heilongjiang Bayi Agriculture University, Daqing* 163316, *Heilongjiang Province, China*)

Abstract: In order to improve the speed, accuracy and efficiency of web data query, we proposed a dynamic web data interactive query algorithm based on user preferences. The user preference model was built to increase the evolutionary individual adaptability of the preference combinations, and the adaptive value was comprehensively calculated. Secondly, in order to prevent data redundancy and duplication, based on interest similarity, query data and duplicate data with high similarity were separated to identify the properties of network data. Finally, the particle swarm optimization algorithm was used to find the optimal interactive query scheme of dynamic web page data. The experimental results show that the quality of the query result set of the proposed algorithm is above 0.95 under the influence of the dataset cardinality, under the influence of the maximum dimension of the query, the quality of the query result set of the proposed algorithm is above 0.96, indicating that the proposed algorithm has short query time, high precision of the result set and strong adaptability.

Keywords: user preference model; dynamic web page data; interactive data query; particle swarm optimization algorithm; spatial dimension

目前，人们的许多信息交流和查询方式都依靠各种软件功能实现，如社会媒体信息动态更新、在

收稿日期：2023-03-23.

第一作者简介：赵红梅(1979—)，女，汉族，硕士，副研究员，从事教育数字化和数据挖掘的研究，E-mail：zhm01230@163.com.

基金项目：黑龙江省教育厅高等教育教学改革研究项目(批准号：SJGY20200508)和大庆市社会科学界联合会项目(批准号：DSGB2020084).

线购物、等待客户在线回复等[1]. 因此，需对海量数据进行分析和查询，以提高系统的性能和效率[2]. 当服务器硬件和数据库的配置固定时[3]，数据量越多，数据查询的速度越慢，且会导致查询卡顿，使网页数据库不能正常工作. 传统的大数据分析和查询方法普遍存在速度较慢的问题，且采用的是不交互的查询方法. 所以，需考虑一种能对动态网页数据进行分析和处理，并能查询到各种交互式数据的算法.

邓斌等[3]提出了一种基于元数据关联特征的交互式数据快速查询方法，通过建立 Map Reduce 编程模型，利用该模型处理元数据，获取元数据关联结果，并建立高维相空间，实现交互式数据快速查询. 周雨佳等[4]使用递归神经网络建立用户的个性化偏好以及用户兴趣的动力学模型，然后通过注意力机制，查询用户的历史行为动态权重，与以往的用户兴趣模型相比，该模型能更多地满足目前的用户需求，最后根据评分显示文档查询结果. 唐运乐等[5]提出了通过动态分布式聚类算法实现大数据的查询，将输入的数据分成若干个子集，以 RR 的形式存于一套计算机节点中，在 Apache Spark 平台上，利用划分和层次动态聚类方法实现数据的分布式聚类，根据 K-近邻查询法，得到查询结果. 虽然上述方法能实现数据查询目标，但存在数据查询精准度较低的问题，会出现结果冗余的情况，影响用户满意度.

为解决上述问题，本文提出一种面向用户偏好的动态网页数据交互式查询算法. 该算法首先构建用户偏好模型，使数据查询结果更符合用户需求，不仅能解决传统方法的数据丢失问题，而且能提高用户满意度；然后通过粒子群优化算法提高搜索能力，实现动态网页数据交互式查询.

1 构建用户偏好模型

用户偏好是指个体对产品或服务作出的合理选择，是使用者在认识和心理上的权衡[6-7]. 构建用户偏好模型可准确掌握用户偏好信息，使数据查询结果更符合用户的需求. 设 M 为一个二元表示的用户偏好模型，$M=(\boldsymbol{I},\boldsymbol{F})$，其中：$\boldsymbol{I}$ 为矩阵，每行表示群体内某一进化个体的基因型；$\boldsymbol{F}$ 为列向量，对应 $\boldsymbol{I}$ 中进化个体的适配值[8]. $\boldsymbol{F}$ 的表达式为

$$\boldsymbol{F}=\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}\begin{pmatrix} f_1 \\ \vdots \\ f_n \end{pmatrix}, \tag{1}$$

其中：a_{ij} 为只能表达特定功能的基因片段，也可被定义为一种特定的代码[9]；f_i 为进化个体的适配值；n 为使用者偏好模式下的演化个体数目，m 为演化个体数目，$i=1,2,\cdots,n$，$j=1,2,\cdots,m$.

为反应使用者偏好，通过用户偏好模型的历史评估数据，选取使用者评估中适配值大于平均分的用户信息，并将进化个体中适配值高于平均值的个体称为偏好个体. 用户偏好模型中的历史评估信息来自用户偏好信息集合. 因此可得组成用户偏好值 $f(a_{ij})$ 的计算公式为

$$f(a_{ij})=\boldsymbol{F}\cdot\sum_{i=1}^{k}\frac{f_i w_j}{l}=\boldsymbol{F}\cdot\frac{1}{l}\sum_{i=1}^{k}f_i w_j, \tag{2}$$

其中：l 为由用户评估的偏好个体总数；k 为由用户评估具有 a_{ij} 偏好的个体数目；w_j 为由 a_{ij} 所对应组成的权重，其中 $\sum_{j=1}^{m}w_j=1$.

基于分析得出的演化个体间的相互关系，提出一种偏好组合. 在演化中，当固定组成部分的组合频率高于临界点[10] α 时，称为偏好组合. 在估计适配值时，需要增加偏好组合的演化个体适应性，并引入偏好组合因子 λ：

$$\lambda=\frac{\mathrm{Freg(Comb)}}{l}\times\frac{\mathrm{Count(Comb)}}{m}, \tag{3}$$

除涉及合并的次数 Freg(Comb)，Freg(Comb)$=1,2,\cdots,l$，Count(Comb)$=2,3,\cdots,m-1$ 外，λ 与组合中的组成成分 Count(Comb)也密切相关，即偏好组合的频率越高，偏好组合中的组成成分越多，λ 值越高.

设 x 为新生成的任一进化个体，α_c 为特定的选择阈值，综合各组成部分的适配值，考虑各部分间的相关性，可得该进化个体 x 的估计适配值为

$$f(x)=\begin{cases}(1+\lambda)\sum_{j=1}^{m}f(a_{ij}), & \alpha\geqslant\alpha_c,\\ \sum_{j=1}^{m}f(a_{ij}), & \alpha<\alpha_c.\end{cases}\tag{4}$$

通过上述分析可知，用户偏好模型建立在每个独立部分的适配值上，可防止发生数据缺陷，最大的适配值个体将会在后代中保存下来，减少数据丢失.

2　动态网页数据相似性度量

通过构建用户偏好模型降低了数据丢失概率，为进一步防止数据冗余和重复，从兴趣相似性出发，得出查询数据和重复数据的相似度，一旦发现有很高的相似性，就能判定对应的数据是冗余数据. 兴趣相似性的计算公式为

$$\mathrm{sim}(v_a,v_b)=\frac{\sum_{k=1}^{m}W_{ak}\times\sum_{k=1}^{m}W_{bk}}{\sqrt{\sum_{k=1}^{m}W_{ak}^2\times\sum_{k=1}^{m}W_{bk}^2}},\tag{5}$$

其中 v_a，v_b 是对 A，B 的兴趣描述，也可以表示为网络特征矢量；W_{ak} 表示网络中的特征字 t_k 与该数据 A 的关联权重；W_{bk} 表示该网络数据的特征字 t_k 与该网络数据 B 的关联权重.

利用式(5)可从复杂的重复数据中提取出与查询数据存在关联相似性的信息，并通过特征矢量表示与网络具有相似性的数据. 矢量的数目反映了特征的数目，两者都由 m 决定，利用式(5)的计算方法可识别出网络数据间的相关性质.

由以上分析可获得所有数据的对应特征关联，当特征没有对应的关联时，即可判断出这些数据与冗余数据的重复属性之间有无对应关系，此时添加一个阈值 ΔL，当阈值 ΔL 小于两个网络属性之间的相似性时，即可对冗余和非冗余数据进行判断. 若数据 A 和 B 所获得的兴趣特征矢量相同，则可推断出 A 和 B 有相同的属性，从而判断出两个属性之间的相似性，并将结果与 ΔL 进行比较，获得两者之间的相似程度，以提高交互式查询的精准度.

3　动态网页数据交互式查询

通过构建用户偏好模型获取用户偏好信息，可提高用户对数据查询结果的满意度，同时，通过度量动态网页数据的相似性，提取属性相同的数据，可提高数据查询的效率与精度. 在此基础上，用粒子群优化算法实现动态网页数据交互式查询方法设计[11].

粒子群优化算法是一种基于鸟类生活习惯的优化算法，每个粒子都有一个相邻的粒子，其速度方向和最优问题的方向相同，但位置不同，所以粒子最优解是最佳的. 设第 i 个粒子的当前速度为 $\boldsymbol{V}_i=(v_{i1},v_{i2},\cdots,v_{id})$，当前位置为 $\boldsymbol{X}_i=(x_{i1},x_{i2},\cdots,x_{id})$，历史最佳位置为 $\boldsymbol{P}_i=(p_{i1},p_{i2},\cdots,p_{id})$，目前粒子群的最佳位置为 $\boldsymbol{P}_g=(p_{g1},p_{g2},\cdots,p_{gd})$，则更新第$(k+1)$个粒子的速度和位置可表示为

$$V_{id}^{k+1}=V_{id}^{k}+c_1r_1(p_{id}-X_{id}^{k})+c_2r_2(p_{gd}-X_{id}^{k}),\tag{6}$$

$$X_{id}^{k+1}=X_{id}^{k}+V_{id}^{k+1},\tag{7}$$

其中：$d=1,2,\cdots,D$，D 为搜寻空间维度；c_1 为个人的认知能力，c_2 为社交能力；r_1，r_2 为一个(0,1)内的随机数.

从标准粒子群算法的工作流程可见，速度与位置对算法的性能有较大影响，但在实际应用中会受很多因素的影响，为提高其的稳定性，对粒子群优化算法进行如下改进.

加入惯性权重系数. 在粒子的飞行速度中增加惯性权值，以实现对粒子的整体和局部寻优，将

式(6)改进为

$$V_{id}^{k+1}=WV_{id}^{k}+c_1r_1(p_{id}-X_{id}^{k})+c_2r_2(p_{gd}-X_{id}^{k}),\tag{8}$$

其中 W 为惯性权重系数.

在初始阶段,粒子群优化算法具有较好的全局寻优能力和局部寻优性能,在后期需要更好的局部搜索能力,故惯性权值 W 的变化形式为

$$W=W-\frac{W_{\max}-W_{\min}}{\mathrm{iter}}(\mathrm{iter}_{\max}-\mathrm{iter}),\tag{9}$$

其中 $W_{\max}$ 和 $W_{\min}$ 为最大、最小惯性权重系数,iter 为目前的粒子群迭代次数,$\mathrm{iter}_{\max}$ 为粒子群的最大迭代次数.

将粒子群优化算法应用于动态网页数据交互式查询问题的求解,寻找最优的动态网页数据交互式查询方案,步骤如下:

1) 引入数据相似性度量结果 $\mathrm{sim}(v_a,v_b)$,构建动态网页数据交互式查询方案的可行解集合;

2) 设迭代次数 $t=0$,并设定最大迭代数 $\mathrm{iter}_{\max}$,参数 c_1,c_2,$W_{\max}$,$W_{\min}$;

3) 针对动态网页交互式查询对象,提出一种基于最少查询成本和最短响应时间的目标函数;

4) 基于动态网页交互式查询对象的约束条件,构造适应度函数;

5) 通过优化每个粒子的适应度函数,得到目前最佳的历史位置;

6) 按式(9)调节权重,并将粒子的速度和位置按式(8)和式(7)进行更新;

7) 设置重复次数 $t=t+1$;

8) 若 $t>\mathrm{iter}_{\max}$,则终止查询;否则,转步骤 6);

9) 最后通过对粒子群优化的定位,得出最终的动态网页交互式查询结果.

4 实验分析

下面通过实验验证本文算法的有效性,将本文算法与递归神经网络算法和动态分布聚类算法进行对比. 实验数据来自数据集 data.gov(http://www.data.gov/),在该数据集中抽取部分用户相关数据形成实验数据集.

通常情况下,用户会根据自己对数据集的理解不断调整自己的偏好,单凭一次查询很难获得高质量的结果集. 因此,根据数据的维数关系,将数据分为正相关、独立和负相关 3 种类型. 其中正相关数据集为同时增加或减少的偏好阈值,独立数据集为训练数据和测试数据分布情况一致的阈值,负相关数据集为增大特定维度和阈值而减小特定维度的阈值. 在查询时,可按用户的交互动态调节偏好阈值.

4.1 交互次数对结果集质量的影响

首先,验证交互次数对结果集质量的影响,结果如图 1 所示. 由图 1 可见,本文方法处理下的 3 个数据集中,结果集质量交互超过 0.85 的顺序依次为正相关数据集、独立数据集、负相关数据集; 4 轮互动后,正相关数据的交互结果集质量为 0.97,后 9 轮正相关的交互结果集合质量均在 0.90 以上,表明本文方法在早期的交互中,可通过调节阈值快速减小结果集合与期望结果集合尺寸之间的偏差,从而快速使结果质量提高到 0.85.

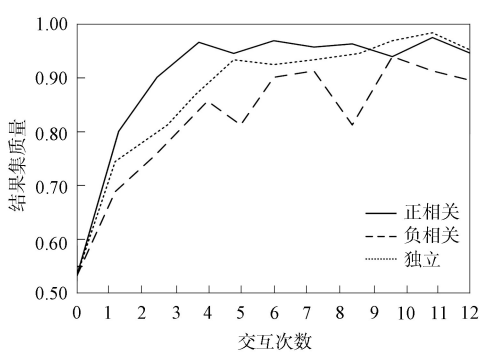

图 1 交互次数对结果集质量的影响

Fig. 1 Effect of interaction times on quality of result set

4.2 数据集基数对算法性能的影响

基数在实验中用于表示查询信息集的大小,实验设置 3 种算法结果集的基数在 200～2 600 内变化,理想结果集不变,数据集基数越大,结果质量越高,则表示查询情况越好. 数据集基数对各算法性能的影响如图 2 所示.

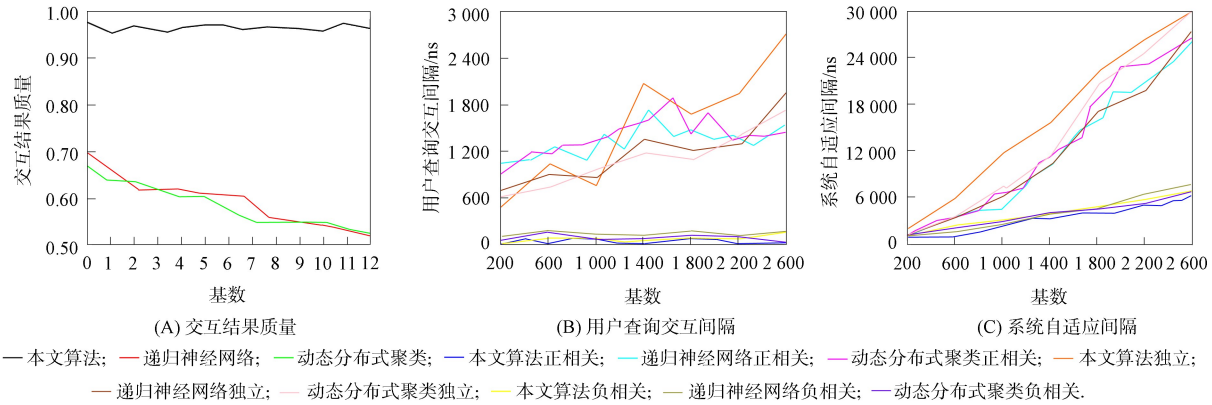

图 2　数据集基数对各算法性能的影响

Fig. 2　Effect of dataset cardinality on performance of each algorithm

由图 2(A)可见：本文算法查询结果集的质量高于递归神经网络算法和动态分布聚类算法；递归神经网络算法和动态分布式聚类算法结果集质量在 0.50～0.70 内，而本文方法查询结果集质量均在 0.95 以上. 同时，随着样本基数的增加，对比算法的结果集质量逐渐降低，对动态网页数据交互式查询效果较差. 由图 2(B)可见，本文算法能有效支持查询用户偏好动态页面数据的变动，交互间隔较短，在 3 种算法中的查询优势明显. 本文算法的平均时间随数据集基数的增加而增加，这主要是因为两种方法每次都对全部数据进行处理，其性能与数据基数之间的关系是线性的，这种情况在图 2(C)中尤其明显.

综上可见，本文算法适用于动态页面的阈值变动，每次执行时需要处理的数据集较少，说明该算法的性能比前两种算法好.

4.3　查询动态网页最大维数对算法性能的影响

选择独立数据集作为查询最大维数的测试对象，查询 1～13 维的变化，结果如图 3 所示. 维度的增大会使结果集包含的信息熵值增大，用户初始选项变多，查询会更困难，需要消耗较多的时间.

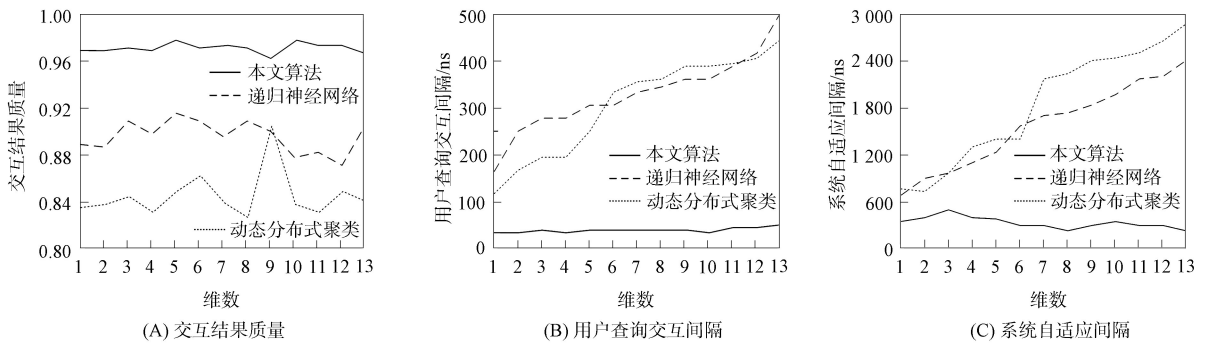

图 3　查询最大维数对各算法性能的影响

Fig. 3　Effect of maximum dimension of query on performance of each algorithm

由图 3 可见，本文算法查询结果集质量在 0.96 以上，而递归神经网络算法在 0.92 以下，动态分布式聚类算法只有在 9 维度时能达到 0.90，本文算法性能不受查询最大维数影响，系统自适应间隔和用户交互间隔数值都比其他两种算法小，说明本文算法的查询效果较好，能有效适应动态网页各种交互式数据类型数据集.

综上所述，针对在数据查询时，由于数据量庞大，传统查询算法存在数据查询精准度较低，且会出现结果冗余情况，无法满足用户需求的问题，本文提出了一种面向用户偏好的动态网页数据交互式查询算法. 首先建立用户偏好模型，防止产生数据丢失问题；然后判断查询数据的相似性，并采用粒子群优化算法增强搜索性能，从而完成动态网页数据交互式查询. 实验结果表明：在数据集基数影响下，本文算法的查询结果集质量均在 0.95 以上；在查询最大维数影响下，本文算法的查询结果集质量

在0.96以上，因此该算法的数据查询效果更好.

参 考 文 献

[1] 赵文涛，张烁. 稀疏数据下基于用户偏好的协同过滤算法 [J]. 重庆邮电大学学报(自然科学版)，2021，33(4)：669-674.（ZHAO W T，ZHANG S. Collaborative Filtering Algorithm Based on User Preference in Sparse Data [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition)，2021，33(4)：669-674.）

[2] 王卫红，曾英杰. 基于聚类和用户偏好的协同过滤推荐算法 [J]. 计算机工程与应用，2020，56(3)：68-73.（WANG W H，ZENG Y J. Collaborative Filtering Recommendation Algorithm Based on Clustering and User Preference [J]. Computer Engineering and Applications，2020，56(3)：68-73.）

[3] 邓斌，陈会平，李凯勇. 基于元数据关联特征的交互式数据快速查询 [J]. 计算机仿真，2021，38(7)：371-375.（DENG B，CHEN H P，LI K Y. Interactive Data Quick Query Based on Metadata Association Characteristics [J]. Computer Simulation，2021，38(7)：371-375.）

[4] 周雨佳，窦志成，葛松玮，等. 基于递归神经网络与注意力机制的动态个性化搜索算法 [J]. 计算机学报，2020，43(5)：812-826.（ZHOU Y J，DOU Z C，GE S W，et al. Dynamic Personalized Search Based on RNN with Attention Mechanism [J]. Chinese Journal of Computers，2020，43(5)：812-826.）

[5] 唐运乐，韦杏琼. 基于动态分布式聚类算法的大数据查询处理方法 [J]. 西南师范大学学报(自然科学版)，2021，46(5)：134-139.（TANG Y L，WEI X Q. Big Data Query Processing Method Based on Dynamic Distributed Clustering Algorithm [J]. Journal of Southwest China Normal University (Natural Science Edition)，2021，46(5)：134-139.）

[6] 朱桂明，宾辰忠，古天龙，等. 基于知识图谱的用户偏好神经建模框架 [J]. 模式识别与人工智能，2019，32(7)：661-668.（ZHU G M，BIN C Z，GU T L，et al. Neural User Preference Modeling Framework Based on Knowledge Graph [J]. Pattern Recognition and Artificial Intelligence，2019，32(7)：661-668.）

[7] SHE W，YANG X Y，TIAN Z，et al. Decentralization Configuration Method of Power Resources Based on User Preference [J]. Automation of Electric Power Systems，2019，32(7)：661-668.

[8] 毛德磊，唐雁. 基于归因理论用户偏好提取的协同过滤算法 [J]. 计算机工程，2019，45(6)：225-229.（MAO D L，TANG Y. Collaborative Filtering Algorithm Based on Attribution Theory for User Preference Extraction [J]. Computer Engineering，2019，45(6)：225-229.）

[9] 张学旺，付康，叶财金，等. 面向医疗区块链的新型轻节点数据查询方法 [J]. 应用科学学报，2022，40(4)：600-610.（ZHANG X W，FU K，YE C J，et al. New Light Node Data Query Method for Medical Blockchains [J]. Journal of Applied Sciences，2022，40(4)：600-610.）

[10] 钱忠胜，涂宇，俞情媛，等. 一种融合用户动态偏好和注意力机制的跨领域推荐方法 [J]. 小型微型计算机系统，2022，43(6)：1335-1344.（QIAN Z S，TU Y，YU Q Y，et al. Approach to Cross-Domain Recommendation Fusing Users' Dynamic Preferences and Attention Mechanism [J]. Journal of Chinese Computer Systems，2022，43(6)：1335-1344.）

[11] 梁明玉，蔡新红，赵咪. 基于改进粒子群算法的光伏系统MPPT控制研究 [J]. 计算机仿真，2021，38(10)：133-139.（LIANG M Y，CAI X H，ZHAO M. Research on MPPT Control of Photovoltaic System Based on Improved Particle Swarm Optimization [J]. Computer Simulation，2021，38(10)：133-139.）

（责任编辑：韩 啸）