

基于语义指向性分析的数据库访问查询优化设计

马晓珺, 刘凌霞

(安阳师范学院 软件学院, 河南 安阳 455000)

摘 要: 在对数据库进行访问查询的过程中, 容易出现数据错误、重复和分散的问题, 传统方法因只支持准确查询, 所以会造成查询精度低下、效率不高, 对此提出一种基于语义指向性分析的数据库访问查询优化设计方法. 通过语义相似度描述语义指向性, 对单词间语义距离的计算进行扩展, 将单词间的语义距离看作是任意两个含义间语义距离的最小值, 依据用户查询和文档间的相似度描述查询相似度, 将查询相似度看作是计算两个文档间的语义相似度. 在语义指向性分析的基础上, 设计数据库访问查询优化设计模型, 详细分析了模型的主要构成, 利用语义信息约束规则, 优化数据库访问查询结果. 仿真实验结果表明, 所提的基于语义指向性分析的数据库访问查询优化设计方法性能稳定, 具有较高的查全率和排序准确性.

关键词: 语义指向性; 数据库; 访问; 查询

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2016)02-0104-05

Database Access Query Optimization Design Based on the Analysis of the Semantic Directivity

MA Xiao-jun, LIU Ling-xia

(School of Software Engineering, Anyang Normal University, Anyang 455000, China)

Abstract: In the process of access to the database query, prone to data errors, repetition and dispersion problem, traditional methods for only support query accurately, cause the query precision is low, the efficiency is not high question, put forward a kind of based on the analysis of the semantic directivity database access query optimization design method, through semantic similarity describes semantic directivity, extensions for the calculation of semantic distance between words, the semantic distance between the words as two words in the semantic distance between any two meanings of the minimum value, on the basis of the degree of similarity between user query and the document description, query similarity query similarity as to calculate the semantic similarity between two documents. In semantic based on the analysis of the directivity, design of database access query optimization design model, detailed analysis of the main components of the model, using semantic information constraint rules, optimize the database access query results. The simulation results show that the proposed based on the analysis of the semantic directivity of database access query optimization design method of stable performance, has higher recall ratio and sorting accuracy.

Key words: semantic directivity; database; access; the query

1 引言

近年来,随着科技和计算机的逐渐发展和广泛应用,大型数据库逐渐成为各大企业计算机应用的基础及核心^[1-2]. 数据规模的迅猛增长、用户数量的不断增加均对数据库提出了更高的要求,因此,研究

一种高效的数据库访问查询方法,提高查询效率具有重要意义,已经成为相关学者研究的重点课题^[3-5].

目前,研究数据库访问查询优化设计的方法有很多,主要包括模糊理论、分布式哈希表方法和属性相关度估计方法等,相关研究也取得了一定的成果,

其中,文献[6]里提出基于网页上下文的数据库访问搜查优化设计,使用网页的实质文本特点达成对数据库的分类标识,完成分类搜寻,但该方法无法合理描述深度信息,造成分类精度以及查全率低下;文献[7]提出一种基于属性相关度的数据库访问查询优化设计方法,依据属性相关度对数据库的大小进行估计,采用词频统计方法实现数据库访问查询优化设计,但该方法存在实现过程复杂的弊端;文献[8]提出一种基于维度的数据库访问查询优化设计方法,该方法从应用维度着手,通过 noSQL 技术进行数据库访问查询优化设计,但该方法无法有效避免数据库访问过程中的数据纠错、消重和整合等问题,造成数据库实体查询性能不佳;文献[9]提出一种基于分布式哈希表 DHT(DistributedHash Table) 的数据库访问查询优化设计方法,对照分布式哈希表实现数据库访问,但该方法只支持精确查询,若未查询到相关文档则会影响到查询优化的召回率;文献[10]基于之用基于子图匹配的数据库访问查询优化设计方法,该方法将子图匹配作为数据库访问查询的基础,通过 RQL 和 SeRQL 实现优化设计,但该方法无法充分支持复杂性访问查询,而且具有运行效率的弊端。

本文提出一种基于语义指向性分析的数据库访问查询优化设计方法,用语义相似度描述语义指向性分析,设计数据库访问查询优化设计模型,通过语义信息约束规则,优化数据库拜访问效果. 仿真实验结果证明,所提方法性能稳定,具备较高的查全率和排序准确性。

2 基于语义指向性分析的数据库访问查询优化设计

2.1 语义指向性分析

语义指向性是指句法构造的某一个成分在语义上与其他成分相配备的可能性,而这样的匹配可能性可通过语义相似度的计算进行描述,因此,本文首先对语义相似度进行计算,从而实现语义指向性分析。

2.1.1 单词间语义相似度的计算

把连接子概念节点 c 与父概念节点 p 边的权值描述成:

$$wt(c, p) = \left[\beta + (1 - \beta) \frac{\bar{E}}{E(p)} \right] \left(\frac{d(p) + 1}{d(p)} \right)^\alpha.$$

$$[IC(c) - IC(p)] \cdot T(c, p) \quad (1)$$

式中, $d(p)$ 用来标识节点在层次构造里的深度;

$E(p)$ 用来表示 p 节点的子节点的个数; $\bar{E}$ 用来表示全部层次构造的平均密度; $T(c, p)$ 用来表示节点之间的关系类型; 参数 α 和 β 分别用来调节节点的深度和部分密度对权值的作用; 概念信息量 IC 能通过下式求出:

$$IC(c) = \log^{-1} P(c) \quad (2)$$

式中, $IC(c)$ 用于表示概念 c 的信息量; $P(c)$ 用来表示概念 c 在一个数据库里产生的概率. 跟着节点产生概率的逐渐增加, 其包含的信息量将逐渐减少。

两个单词间的语义距离可通过下式求出:

$$\text{Dist}(w_1, w_2) = \sum wt(c, \text{parent}(c)) \quad (3)$$

式中, $c_1 = \text{sen}(w_1)$; $c_2 = \text{sen}(w_2)$; $\text{path}(c_1, c_2)$ 用于描述从 c_1 至 c_2 最短途径里全部节点的集合; $L_{\text{Super}}(c_1, c_2)$ 用来表示概念 c_1, c_2 最底层的上位概念; $\text{parent}(c)$ 用来表示概念 c 的父概念, 但上述方法只适合单词只有一个含义的状况. 本文对单词间语义间隔的计算实行扩展, 自理单词包含多个含义的状况. 数据库里两个概念节点之间的间隔即沿最短途径上全部边的权值之和, 则概念节点间的距离可描述成:

$$\begin{aligned} \text{Dist}(c_1, c_2) &= IC(c_1) + IC(c_2) - \\ &2 \times IC(L_{\text{Super}}(c_1, c_2)) \end{aligned} \quad (4)$$

如果某个单词留有许多意义, 则单词间的语义间隔就是两个单词里任何两个含义间语义距离的最小值, 也就是:

$$\text{Dist}(w_1, w_2) = \min[\text{Dist}(c_1, c_2)] \quad (5)$$

式中, $\text{sen}(w)$ 用于描述单词 w 整个含义的集合。

2.1.2 查询相似度的计算

本文对查询相似度使用用户查询与文档间的相似度实行表示, 把查询相似度当作是评价两个文档间的语义相似度. 加入把文档当作是概念的集合, 那么文档的查询相似度能经过包含有关概念的语义相似度实行代替, 则有:

$$\begin{aligned} \text{Dist}(q, d) &= \text{Dist}(X_1 \wedge X_2 \cdots \wedge X_k, Y_1 \wedge Y_2 \cdots \wedge \\ Y_m) &= \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^m f_j \cdot \text{Dist}(X_i, Y_j) \end{aligned} \quad (6)$$

式中, X_i 用来表示用户给出的查询重点词, Y_j 用来标识文档 d 的简单 RDF 概念列表中的单词, 简易

RDF 概念列表是文档的表示模式; f_j 用来表示概念 Y_j 在文档 d 里产生的次数; m 用来表示文档简单 RDF 概念列表的元组数量; n 用来表示计算两个文档里概念语义间隔的个数,其实际上是标准化文档的间隔。

那么查询相似度可描述成:

$$\text{Sim}(q, d) = \begin{cases} \frac{1}{1 + \text{Dist}(q, d)}, & \text{if}(l_q = l_d) \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

因为简易 RDF 概念列表中属性能够对文档主题进行标识,不属于同一主题的文档其相似度是 0,所以,可以在两个文档属于同一主题的情况下对二者的语义相似度进行计算,从而减少计算过程,提高查询效率。语义相似度越高,句法结构的某一成分在语义上和其他成分(一个或几个)相匹配的可能性则越高,依据上述语义指向性分析对数据库方位查询实行优化设计。

2.2 模型设计及算法实现

在语义指向性加系的根本上,设计数据库访问查询优化模型,模型重点包含文本配备模型、语义指向性解析模型与分组统计模型,使用语义信息对规则实行约束,优化数据库访问查询后果,数据库访问查询优化设计模型用图 1 进行描述。

利用语义指向性分析对数据资源的选择、消重和数据整合进行调控。假设数据库实体集为 $x = (x_1, x_2, \dots, x_m)^T$,通过语义指向性解析和静态解析文本相似集的线性掺杂,对语义信息约束规则实行表示,改进逼近数据库访问查询后果,在语义指向性解析的根本上,获得语义信息约束规律:

$$x = Hs = \sum_{j=1}^n h_j s_j \quad (8)$$

式中, $H = [h_1, h_2, \dots, h_n]$ 用于描述 $m \times n$ 阶满秩源信号混合矩阵;则数据库访问查询优化设计模型输出模式可转换成下述形式:

$$\begin{bmatrix} x_1(t) \\ \dots \\ x_m(t) \end{bmatrix} = \begin{bmatrix} h_{11} \dots h_{1n} \\ \dots \\ h_{m1} \dots h_{mn} \end{bmatrix} \begin{bmatrix} s_1(t) \\ \dots \\ s_2(t) \end{bmatrix} \quad (9)$$

在相同实体属性特征范围内,采用本文数据库访问查询优化设计模型对语义表象之间的指向性和相似度进行编辑,通过属性选择算法,在 $s = Ax$ 的情

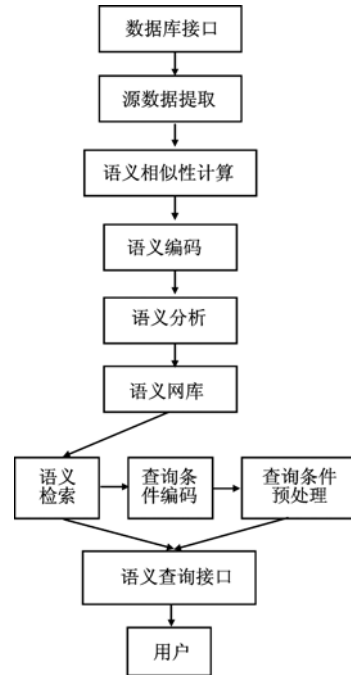


图1 数据库访问查询优化设计模型

况下,对矩阵的逆矩阵实行计算,获得组合特点 s 及属性级的相似度。利用前文的语义指向性分析达成数据库的原始文本相似集的估算,获得辅助信息库。源数据库里语义特点 x 经分离矩阵 W 实行逆变换后,生成的输出特点可描述成:

$$y = [y_1, y_2, \dots, y_n]^T \quad (10)$$

综上所述,通过语义指向性分析的语义分析模式可描述成:

$$y(t) = Wx(t)\text{Sim}(q, d) = WHs(t)\text{Sim}(q, d) = Gs(t)\text{Sim}(q, d) \quad (11)$$

式中, G 用来表示整个传输矩阵。经过语义信息约束规则,得到数据库访问查询后果,经过自适应调整 $G = I$ (I 是 $n \times n$ 阶单位矩阵),获取 $y(t) = s(t)$,从而实现对源信号的分离,以适应数据库动态关性的要求。使用语义指向性分析组建关联知识库,达成数据库访问查询进程里的纠错、消重及整合治理。图 2 是语义指向性数据库访问查询优化设计的原理图。

3 仿真实验分析

为了证明本文提出的基于语义指向性解析的数据库访问查询优化算法的有用性,须要实行有关的验证解析。实验将 DHT 方法作为对比进行分析。

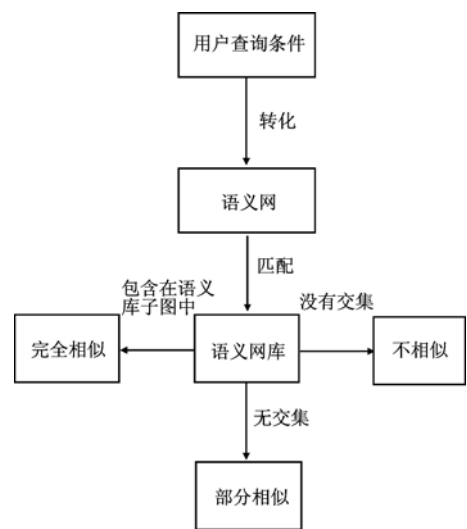


图 2 语义自相关特征数据库实体识别匹配原理图

3.1 稳定性分析

表 1 描述的是本文方法和 DHT 方法语义相似的查询样本集合. 分析表 1 可以看出,采用本文方法得到的查询之间的语义相似度相差不大,且一直高于 DHT 方法,而采用 DHT 方法得到的查询之间的语义相似度波动较大,说明本文方法具有较高的稳定性.

表 1 相似度评估的鲁棒性

Query Q		本文方法	DHT 方法
商品名 = C 语言程序设计	商品名 = C + + 语言算法与程序设计	0.83	0.73
	商品名 = 全国计算机等级考试:二级 C 语言	0.75	0.65
	商品名 = 数据结构	0.58	0.41
商品名 = 三国演义	译著者 = 罗贯中	0.94	0.82
	商品名 = 水浒传	0.51	0.44
	商品名 = 三国志	0.22	0.2

3.2 响应时间分析

表 2 描述的是在 200 Mb/s 以太网环境中,在数据库规模从 50 ~ 400 的情况下,采用本文方法和 DHT 方法进行数据库访问查询的响应时间比较结果.

分析表 2 可以看出,对大规模数据库进行访问查询的过程中,本文方法显著提高了响应速度,同时,在数据库规模逐渐增加的情况下,采用本文方法的响应时间增长幅度也明显低于 DHT 方法,验证了本文方法的有效性.

表 2 查询响应时间对比数据

数据规模	本文方法	DHT 方法
50	1.3	16.5
100	1.6	17.4
150	1.1	21.1
200	0.9	15.9
250	2.2	18.4
300	1.8	14.8
350	1.5	22.5
400	1.4	23.4

3.3 查全率分析

查全率就是查询结果中的相关元组数与数据集中的相关元组总数之比,采用本文方法和 DHT 方法的查全率比较结果用图 3 进行描述.

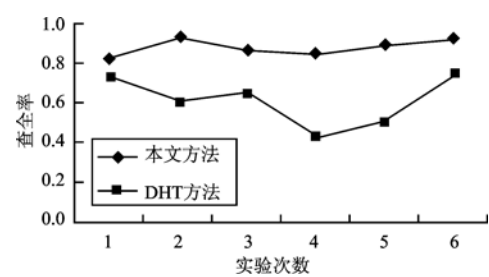


图 3 两种方法查全率比较结果

分析图 3 可以看出,本文方法的查全率曲线明显优于 DHT 方法,同时本文方法的查全率一直保持在 0.8 以上,相对比较平稳,说明本文方法不仅具有很高的性能,而且稳定性也较优.

3.4 排序准确性分析

两个集合间一样的元组数增加,排序方法的正确性就越高,用户觉得最为有关的后果被排到了前面. 本文算法和 DHT 方法的排序准确性比较结果用表 3 进行描述.

表 3 两种方法排序准确性比较结果

序号	本文方法	DHT 方法
1	0.88	0.71
2	0.79	0.62
3	0.86	0.59
4	0.82	0.68
5	0.91	0.66
6	0.85	0.73
7	0.76	0.78

分析表 3 可以看出,采用本文方法的排序准确性均在 0.75 以上,明显高于 DHT 方法,且一直高于 DHT 方法,说明本文方法能够较好地满足用户偏好.

4 结束语

本文提出一种基于语义指向性分析的数据库访问查询优化设计方法,通过语义相似度描述语义指向性,对单词间语义距离的计算进行扩展,把单词间的语义间隔当作是两个单词里任何两个含义间语义间隔的最小值,依据用户查询与文档间的相似度表示查询相似度,父查询相似度当作是计算两个文档间的语义相似度.在语义指向性分析的基础上,设计数据库访问查询优化设计模型,详细分析了模型的主要构成,利用语义信息对规则进行约束,优化数据库访问查询结果.仿真实验结果表明,所提出的基于语义指向性分析的数据库访问查询优化设计方法性能稳定,具有较高的查全率和排序准确性.

参考文献:

- [1] 万长林,徐德明.基于语义的数据库关键词检索系统[J].计算机与现代化,2014,26(10):37-41.
- [2] 夏海峰,陈军华.基于语义相似度计算的 DeepWeb 数据库查询[J].微型机与应用,2014,33(8):64-67.

- [3] 王庆荣.基于语义异构数据库的概念格分布式构造算法研究[J].兰州交通大学学报,2013,32(3):74-77.
- [4] 于彤,刘静,刘丽红,等.面向中药数据库的语义集成框架[J].中国数字医学,2015,10(1):78-80.
- [5] 涂中群,端木艺.基于语义的个性化关联检索模型构建研究——以楚辞研究数据库的实现为例[J].图书情报工作,2013,57(14):121-124.
- [6] 王林彬,黎建辉,沈志宏.基于 NoSQL 的 RDF 数据存储与查询技术综述[J].计算机应用研究,2015,31(5):1281-1286.
- [7] 刘伟.医疗数据库松弛响应关联维分析及差异查询[J].科技通报,2014,29(6):116-118.
- [8] 吴珺,谢树泉,王春枝.基于语义的隧联网数据整合方法[J].智能系统学报,2014,9(6):704-708.
- [9] 唐晓婷,吴爱华,曾卫明.不一致数据库中基于用户语义模板的评论可信度计算[J].燕山大学学报,2014,38(6):523-531.
- [10] 臧强,程立.基于 ASP 技术和 SQL 数据库的成绩查询系统的设计与实现[J].电子设计工程,2014,22(3):45-47.

作者简介:

马晓珺 女,(1980-),硕士,讲师.研究方向为信息存储检索和图形图像处理. E-mail: maxj@aynu.edu.cn
刘凌霄 女,(1977-),硕士研究生,副教授.研究方向为图形图像处理、智能信息控制.