基于概率统计及查询质量的数据源选择策略

张春海 刘 群 李 华
(中国海洋大学计算机科学系 青岛266071)

摘 要 数据源选择策略对提高异构数据集成效率具有重要意义。综合考虑概率统计信息和数据源相对执行质量,提出一种新的数据源选择策略。概率统计信息通过分析查询日志得到,包括针对频繁主题类的数据源覆盖度和数据源集合重叠度。数据源相对执行质量包括查询价格、执行时延、数据源可靠性和用户满意度。给出基于以上标准的数据源选择贪心算法,并通过实验原型验证算法有效性。

关键词 异构数据集成,重叠度,覆盖度,相对执行质量

Data Source Selection Strategy Based on Probability Statistics and Quality Measurement

ZHANG Chun-Hai LIU Qun LI Hua
(Department of Computer Science, Ocean University of China, Qingdao 266071)

Abstract Data source selection strategy plays an important role in improving the efficiency and effectiveness of heterogeneous data integration. This paper proposes a greedy algorithm to select proper data sources by combining probabilistic knowledge in mediated system and qualitative information about quality of data sources. Two kinds of probabilistic information of frequent topic classes are considered, which are coverage of each data source and overlap degree between different data sources. Quality of data sources includes query price, response time, reliability and user satisfaction. Finally, a preliminary experiment is carried out to evaluate the efficiency of the proposed algorithm.

Keywords Heterogeneous data integration, Overlapping probability, Coverage probability, Quality of data sources

1 引言

基于中间件的异构数据集成方法针对用户查询,给出基于中间件数据模式的查询转换,将用户查询变换成对多个异构数据源的查询,最后集成各个数据源的查询结果,返回给用户[1]。中间件的首要任务是设计一个公共数据模式,为用户提供全局统一的视图;其次是进行全局执行规划,利用概率统计信息,计算数据源的相关程度,由此确定以何种次序存取这些数据源的数据[2~4]。本文通过改进评估查询效率的标准[2],把查询准确性和查询质量相结合,确定数据源的存取顺序,以期提高异构数据集成效率。

为提高查询准确性,通过分析查询日志,根据查询频率确定频繁主题类、数据源覆盖度和数据源集合重叠度,选择以最大可能包含查询结果的数据源。

对于查询质量,考虑集成系统中不同数据源的相对执行质量[5,6],把查询价格、执行时延、数据源可靠性、用户满意度作为数据源质量优劣的决定因素。

2 异构数据集成系统基本框架

基于中间件的异构数据集成系统由数据源层、中间件和用户界面三部分构成。中间件包括两个全局模式,即数据集成模式和抽象模式。

数据集成模式实现各个异构数据源数据模式的映射。现有的模式匹配方法包括:实例层匹配、模式层匹配、元素级匹配、结构级匹配、基于约束的方法等[7]。

抽象模式定义主题类概率模型,通过分析查询日志,计算不同数据源和给定查询的相关程度。由于时空限制,不可能统计所有用户可能提交的查询的概率,因此需要把用户查询映射到相应的主题类,计算不同数据源对于主题类的覆盖度,由此确定数据源的存取顺序。最后向选定的数据源提交原始查询。图1给出异构数据集成的基本框架。

3 概率统计信息描述

3.1 利用概率统计信息实现查询优化

中间件的一个目标就是选择 k 个数据源,最大化概率 $P(S_1 \vee \cdots \vee S_k|Q)$,其中 $P(S|Q)$ 表示随机选取一个满足查询 Q 的数据对象,出现在数据源 S 中的概率。由于用户提交的查询并非服从均匀分布,中间件利用查询日志,进行聚类分析,把分散的查询抽象到主题类,通过计算数据源针对主题类的覆盖度,选择合适的数据源。

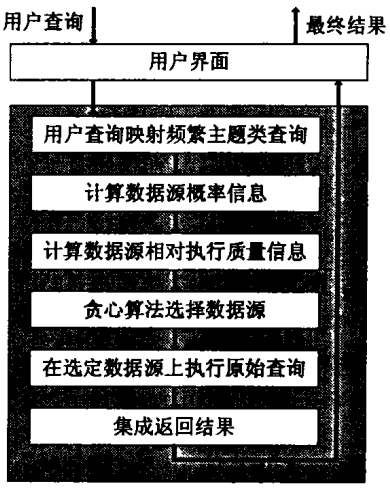

图1 异构数据集成基本框架

主题类可以通过文[4]提出的主题类概念分层结构得到。

张春海 副教授,硕士生导师,主要研究方向是数据库理论与数据库系统、异构数据集成。刘 群 硕士研究生,主要研究方向是异构数据集成。李 华 副教授,主要研究方向是数据库理论及应用。

利用对象的可分类属性(即具有较高区分度的属性),建立相应的属性值层次树。多个可分类属性的属性值层次树相互连接,形成主题类概念分层结构,由此一对一确定查询 Q 的主题类 C。

3.2 查询日志与频繁主题类

查询日志记录所有提交给中间件的查询请求,包含查询 Q、所属主题类 C、查询出现频率、返回结果数和不同数据源的相应覆盖度等。根据查询日志,可以得到查询 Q 的概率 $P(Q)=\frac{Fr_Q}{Fr}$,其中 Fr_Q 表示查询 Q 在查询日志中出现的次数,Fr 表示所有查询次数之和。由此得到主题类 C 的概率 $P(C)=\sum_{Q\in C}P(Q)$。若 $P(C)$ 大于给定的最小阈值 min_fre,则 C 为频繁主题类。中间件仅保留不同数据源集合所包含的和频繁主题类相关的概率统计信息。

给定查询 Q,首先利用主题类概念分层结构,把 Q 映射到相应的频繁主题类。方法如下:

第一步:确定查询 Q 的主题类 C,第二步:若 C 为频繁主题类,则返回 C,否则在主题类概念分层结构中,从 C 向上寻找距离最近并且是频繁主题类的祖先节点,第三步:若第二步得到同一层次上多个祖先节点(如 C_x 和 C_y),且均为频繁主题类,则通过单一数据源的平均主题类覆盖度进行取舍。

3.3 利用数据源重叠度选择数据源

假设已经存取数据源 $S_1\cdots S_{i-1}$ 中数据,由于数据源彼此重叠,对于数据源 S_i,必须计算 S_i 中与 $S_1\cdots S_{i-1}$ 不重叠数据的概率[2,4]。

$$P(\neg S_1\wedge\cdots\wedge\neg S_{i-1}\wedge S_i|C)=P(S_i|C)-\sum_{1\leqslant j\leqslant i-1}P(S_j\wedge S_i|C)+\sum_{1\leqslant j<k\leqslant i-1}P(S_j\wedge S_k\wedge S_i|C)+\cdots+(-1)^{i-1}P(S_1\wedge\cdots\wedge S_{i-1}\wedge S_i|C)$$

其中,

$$P(S_1\wedge\cdots\wedge S_{i-1}|C)=\frac{\sum_{Q\in C}P(S_1\wedge\cdots\wedge S_{i-1}|Q)P(Q)}{P(C)}$$

和

$$P(S_1\wedge\cdots\wedge S_{i-1}|Q)=\frac{N_{Q(S_1\wedge\cdots\wedge S_{i-1})}}{N_Q}$$

是来自中间件的类源概率统计信息,其中 N_Q 表示各数据源返回的满足查询 Q 的结果个数之和,$N_Q(S_1\wedge\cdots\wedge S_{i-1})$ 表示查询结果中同时属于数据源 $S_1\cdots S_{i-1}$ 的结果个数。

4 数据源相对执行质量信息

4.1 数据源执行质量的定义

对于数据源 S,考虑如下四个非功能性属性作为决定数据源质量优劣的因素[5,6]。

(1)查询价格 $Price(S)$:表示用户查询数据源 S,需要支付的费用。查询价格由数据源提供者确定。

(2)执行时延 $Time(S)$:表示向数据源 S 提交查询和得到查询结果之间的预期时延。可以表示为:

$$Time(S)=Exec(S)+Trans(S)$$

其中 $Exec(S)$ 表示查询在数据源 S 上的执行时间,$Trans(S)$ 表示查询结果在网络上的传输时间。利用查询日志记录的信息进行计算:$Time(S)=\frac{\sum_{i=1}^{n}\overline{time_i}}{n}$,其中 $\overline{time_i}$ 表示第 i 次请求查询数据源 S 需要的执行时延,n 表示 S 总共被请求查询的次数。

(3)数据源可靠性 $Rel(S)$:表示提交的查询在最长等待时间内被正确响应的概率。$Rel(S)=N(S)/n$,其中 $N(S)$ 表示正确响应的次数,n 表示 S 总共被请求查询的次数,有 $0\leqslant Rel(S)\leqslant 1$。

(4)用户满意度 $Satis(S)$:表示数据源的用户满意程度,包括数据源中数据的正确性、实时性和完整性,是数据源数据的综合评价,通过用户反馈机制得到。$Satis(S)=\frac{\sum_{i=1}^{n}\overline{satis_i}}{n}$,其中 $\overline{satis_i}$ 表示第 i 个用户对数据源 S 的评价,且 $0\leqslant\overline{satis_i}\leqslant 1$,$n$ 表示用户对 S 的总共评价次数,有 $0\leqslant Satis(S)\leqslant 1$。

考虑以上四个标准,数据源 S 的执行质量向量可以表示为:

$$\overrightarrow{q(S)}=(Price(S),Time(S),Rel(S),Satis(S))$$

4.2 数据源相对执行质量与规范化处理

假设数据集成系统共包含 n 个数据源 $S_1\cdots S_{i-1}$。对于数据源 S_i,它的执行质量向量 $\overrightarrow{q(S_i)}=(Price(S_i),Time(S_i),Rel(S_i),Satis(S_i))$,记作 $\overrightarrow{q(S_i)}=(q_{i,1},q_{i,2},q_{i,3},q_{i,4})$。整个系统 n 个数据源的执行质量向量构成如下 $n\times 4$ 维矩阵:

$$\begin{pmatrix}q(S_1)\\ \vdots\\ q(S_n)\end{pmatrix}=\begin{pmatrix}q_{1,1}&q_{1,2}&q_{1,3}&q_{1,4}\\ \vdots&\vdots&\vdots&\vdots\\ q_{n,1}&q_{n,2}&q_{n,3}&q_{n,4}\end{pmatrix},$$

设 $q_j^{max}=\mathrm{Max}(q_{i,j}),1\leqslant i\leqslant n$,$q_j^{min}=\mathrm{Min}(q_{i,j}),1\leqslant i\leqslant n$,利用以下公式进行矩阵列向量元素的规范化处理,把数据源执行质量向量中四个分量的值转换为集成系统中相对数值。

当 $j=1,2$ 由于分量 $q_{i,1}$ 表示的查询价格和分量 $q_{i,2}$ 表示的执行时延的数值与数据源的执行质量成反比,故查询价格、执行时延可以规范化为:

$$v_{i,j}=\begin{cases}\frac{q_j^{max}-q_{i,j}}{q_j^{max}-q_j^{min}} & \text{若 } q_j^{max}-q_j^{min}\neq 0;\\ 1 & \text{若 } q_j^{max}-q_j^{min}=0。\end{cases}$$

当 $j=3,4$,由于分量 $q_{i,3}$ 表示的数据源可靠性和分量 $q_{i,4}$ 表示的用户满意度的数值与数据源的执行质量成正比,故数据源可靠性、用户满意度可以规范化为:

$$v_{i,j}=\begin{cases}\frac{q_{i,j}-q_j^{min}}{q_j^{max}-q_j^{min}} & \text{若 } q_j^{max}-q_j^{min}\neq 0;\\ 1 & \text{若 } q_j^{max}-q_j^{min}=0。\end{cases}$$

通过以上规范化处理,得到 $0\leqslant v_{i,j}\leqslant 1$,其中 $j=1\cdots 4,1\leqslant i\leqslant n$。用向量 $\vec{V}_i=(v_{i,1},v_{i,2},v_{i,3},v_{i,4})$ 表示数据源 S_i 在集成系统中的规范化执行质量向量。不同的客户查询,对数据源不同的非功能性属性有不同的要求。用向量 $\vec{W}=(w_1,w_2,w_3,w_4)$ 表示四个非功能性属性的权,满足 $0\leqslant w_i\leqslant 1$,且 $\sum_{i=1}^{4}w_i=1$。由此,数据源 S_i 的相对执行质量可以计算为:

$$Cost(S_i)=\frac{\vec{V}_i\cdot\vec{W}}{|\vec{V}_i||\vec{W}|}$$

其中

$$\vec{V}_i*\vec{W}=v_{i,1}*w_1+v_{i,2}*w_2+v_{i,3}*w_3+v_{i,4}*w_4$$

$$|\vec{V}_i|=\sqrt{v_{i,1}^2+v_{i,2}^2+v_{i,3}^3+v_{i,4}^2}$$

$$|\vec{W}|=\sqrt{w_1^2+w_2^2+w_3^2+w_4^2}$$

通过上述计算,数据源 S_i 的相对执行质量 $Cost(S_i)$ 被限定在 $[0,1]$。$Cost(S_i)$ 值越趋近于1,说明数据源 S_i 的相对执行质量越高。

5 基于概率信息及相对执行质量的数据源选择算法

5.1 启发式选择贪心算法

考虑网络带宽等限定条件,假设中间件可以并发存取 k

个数据源，并确定数据源存取顺序，即最先选择数据源 $S_1\cdots S_k$，使得 $S_1\cdots S_k$ 能够以最高效率完成查询任务。中间件在选择数据源实现数据集成的过程中，必须同时考虑查询准确性和查询质量两个效率指标，即同时使用概率统计信息和数据源的执行质量作为数据源选择的标准，平衡两者关系，以保证综合效率达到最高。

对于查询准确性，利用查询日志，确定频繁主题类，计算概率统计信息，保证所选择的数据源以最大可能性包含查询结果，即选择 k 个数据源，使得概率 $P(S_1\vee\cdots\vee S_k|Q)$ 取值最大。同时考虑各个数据源的执行质量信息。

若要进行最优数据源的选择，系统需要考虑所有 k 个数据源组成的子集的效率，需进行 $O(n^k)$ 次计算。为了避免运行时进行如此多的计算，启发式选择贪心算法可以很好地实现数据集成过程中数据源的选择。算法允许用户与集成系统交互，根据实际情况，给出数据源概率信息和数据源执行质量特定的权值，同时给出用于确定频繁主题类的阈值 min_fre。

输入：用户查询 Q，数据源 $S_1\cdots S_n$，并发存取数据源个数 k，阈值 min_fre，数据源概率信息的权 w_1 和数据源执行质量的权 w_2，满足 $w_1+w_2=1$。

输出：以 $S\_solution$ 表示的选取的 k 个数据源 $S_1\cdots S_k$。

Pro PQ-greedy-select($Q,S_1\cdots S_n,k,min\_fre,w_1,w_2$)
$S=\{S_1,\cdots,S_n\}$
$S\_solution=\{\}$
根据阈值 min_fre 确定查询 Q 的频繁主题类 C
while *not empty*(S) and $|S\_solution|<k$ do
  for *each* $s\in S$
    $pr=P(s\wedge\wedge_{s_i\in S\_solution}S_i|C)$
    $pc=Cost(s)$
    $ps=\dfrac{pr\times w_1+pc\times w_2}{\sqrt{pr^2+pc^2}\cdot\sqrt{w_1^2+w_2^2}}$
  end for
  $S_{max}=S$ 中 Ps 值最大的数据源
  $S\_solution=S\_solution\cup S_{max}$
  $S=S-s_{max}$
end while
return $S\_solution$
End PQ−greedy−select

5.2 算法有效性验证

为了验证 PQ-greedy-select 算法的有效性，我们进行了原型实验。10000条计算机类图书信息分布在12个独立数据源中，并发存取的数据源个数 k 从1变化到12，同时提供最初模拟查询日志。对于数据源执行质量，主要考虑数据源的执行时延和数据源可靠性。作为比较，实验通过时间复杂度为 $O(n^k)$ 的算法，计算出已知条件下，不同 k 对应的最优数据源集合。图2比较了基于概率统计信息和查询质量的选择贪心算法(PQ-greedy-select 算法)、仅考虑概率统计信息的选择贪心算法(P-greedy-select 算法)和随机选择算法(Random-select 算法)平均返回数据源的精确度。横轴表示频繁主题类的最小阈值 min_fre，纵轴表示结果精确度，$precison=\frac{1}{12}\sum_{k=1}^{12}\frac{|S_{selected}\cap S_{optimal}|}{|S_{selected}|}$，其中 $S_{optimal}$ 表示相应 k 下最优数据源集合，$S_{selected}$ 表示通过相应算法选择的数据源集合，$|S_{selected}\cap S_{optimal}|$ 表示算法选择的数据源集合与最优数据源集合公共的数据源个数。

图3表示并发存取2个数据源以及不同的最小阈值 min_fre 条件下，三个算法返回查询结果个数比较。

通过分析图2、图3可知，基于概率统计信息和查询质量的选择贪心算法(PQ-greedy-select 算法)的效率明显优于随机选择算法，也优于仅考虑概率统计信息的选择贪心算法(P-greedy-select 算法)。总之，PQ-greedy-select 算法在数据集成过程中能够准确有效地选择合适的数据源。

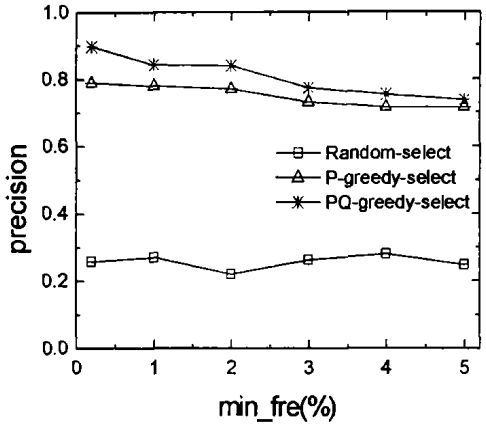

图2 不同最小阈值 min_fre 下，算法平均返回数据源的精确度比较

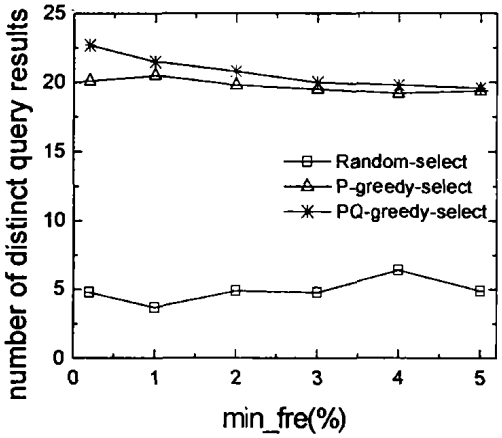

图3 不同最小阈值 min_fre 下，算法返回查询结果个数比较(并发存取2个数据源)

结束语 在基于中间件的异构数据集成系统中，由于网络带宽等条件限制，必须决定以何种顺序查询数据源。为提高查询效率，本文综合考虑查询准确性和查询质量，提出一种新的实现异构数据集成的数据源选择策略，通过分析查询日志，考虑包括频繁主题类、数据源覆盖度和数据源集合重叠度在内的概率统计信息，提高查询准确性。查询质量考虑数据源相对执行质量。最后给出启发式选择贪心算法，并通过实验验证算法有效性。

参 考 文 献

1 Adali S, Candan K. Query Caching and Optimization in Distributed Mediator Systems. In: Proc. SIGMOD'96, Montreal, Canada, June 1996
2 Florescu D, Koller D, Levy A. Using Probabilistic Information in Data Integration. In: Proc. of VLDB'97, Athens, Greece e, August 1997
3 Nie Z, Kambh S. Joint Optimization of Cost and Coverage of Query Plans in Data Integration. In: Proc. of the 10th Intl. Conf. on Information and Knowledge Management, Atlanta, Georgia, Nov. 2001
4 Nie Z, Kambh S. BibFinder/StatMiner: Effectively Mining and Using Coverage and Overlap Statistics in Data Integration. In: Proc. of VLDB'03, Berlin, Germany, Sep. 2003
5 Naumann F, Leser U. Quality-driven Integration of Heterogeneous Information Systems. In: Proc. of VLDB'99, Edinburgh, Scotland, Sep. 1999
6 Zeng L, Enatallah B. Quality Driven Web Services Composition. WWW2003, Budapest, Hungary, May. 2003
7 Rahm E, Bernstein P. A Survey of Approaches to Automatic Schema Matching. VLDB Journal, 2001, 10(4): 34～350