计算机科学2002Vol.29№.7

语义查询优化技术研究综述

Semantic Query Optimization Research: A Survey

何增有　邓胜春　徐晓飞　宋玉福

（哈尔滨工业大学计算机科学与工程系　哈尔滨150001）

Abstract　Semantic query optimization (SQO) is comparatively a recent approach for the transformation of given query into equivalent alternative queries using matching rules in order to select an optimal query based on the costs of executing alternative queries. With the development of database technology and new applications of information systems, data become more complex and heterogeneous. In such situation, traditional query optimization technology is not very suitable. Therefore, since the early 1980's researchers regarded SQO as a promising technique. But no extensive implementations of SQO exist today due to some key problems have not been solved efficiently. In this paper, we survey the research in the area of SQO, point out the process of current research and future directions.

Keywords　Semantic query optimization, Integrity constraint, Semantic rule

1. 引言

传统的查询优化器利用语法变换对查询进行优化，从生成的查询计划中选择一个具有最小代价的执行计划。然而，随着数据库技术和网络技术的发展，尤其是在异构数据库环境下和面向对象的数据库中，处理的对象结构更为复杂，传统的查询优化器显得力不从心。语义查询优化利用数据库上的语义规则将一个查询变为一个语义等价且更加高效地查询，以此来弥补传统查询优化技术的不足。虽然，语义查询优化能够产生较好的优化效果，但必须有效地解决以下一些问题：

·语义规则的内容和表示形式。由于数据库上的完整性约束是多种多样的，这其中包括函数依赖、包含依赖、值域信息、关联规则等等，其关键在于如何有效地表示这些完整性约束以及语义规则应包含哪些完整性约束，使其对语义查询优化提供强有力的支持是有待研究的问题。

·语义规则的自动获取。语义规则是进行语义查询优化的基础，人工地获取它们是低效且不现实的。因此，怎样自动地获取高质量有益于查询优化的语义规则，是影响语义查询优化技术深入应用的瓶颈之一。只有很好地解决了这个问题，才有可能对语义查询优化技术进行更深层次的应用。

·语义规则的维护。绝大部分对查询优化有意义的语义规则都是与数据库的状态相关的，即并非在所有的数据库状态下都成立。因此，在数据库的状态发生变化时，有些规则就会不成立，如果不能有效地处理这些在变化后的数据库状态中不再成立的规则，会影响到查询优化的质量与正确性。于是，如何有效地对这些规则进行维护，同样是需要深入研究的问题。

·查询优化问题。这其中包括可应用规则的选取、等价优化查询的选取以及优化与执行的代价模型，这是语义查询优化技术的核心问题。

在下面的几节中，我们将对上述几个问题的研究现状分别加以介绍，最后对将来进一步的研究工作加以展望。

2. 语义规则的内容和表现形式

由于数据库上的完整性约束是多种多样的，它定义了数据库内在的语义信息[1~3]，对于不同的查询、不同的约束条件所产生的优化效果是不同的，因此，首先我们给出完整性约束的几种不同分类方式，然后讨论语义规则应包含哪些约束以及它们会对语义查询优化的质量产生的影响，这方面在以前的研究中并未受到足够的重视。

按照完整性约束的表征方式，可以将完整性约束分为隐含约束、固有约束以及外在约束。隐含约束是在数据模型中隐式表征的，例如关系模型中的函数依赖；固有约束是一类由数据模型本身决定的约束，无须在模式中指定，关系模型中固有约束的一个例子是属性值的原子性和不可再分性；外在约束是在数据上显示指定的约束，一个属性的值域或属性值之间的复杂的关系便属于此类约束。

另一种分类方式是基于数据库的状态变化对约束的影响来划分的。如果约束在数据库的任意状态都成立，此类约束称之为永久约束；另一类约束与数据库的状态是相关的，此类约束是在数据库的特定状态下成立的，当数据库的状态发生改变时，约束可能不再成立，我们将此类约束称之为状态约束。

不同类型的约束对语义查询优化效果产生的影响是不同的，因此，我们应当对这个问题加以研究，从而有效地组织语义规则的构成，使之具有高可用性。由于约束对语义查询优化的贡献是与特定的查询相关的，因此我们在讨论时认为被检测的约束被给定查询的约束条件所蕴含，此时，待测约束便可加入到查询的约束条件中，变形后的查询会具有与初始查询不同的执行代价（即可能代价变小，这是我们进行语义查询优化研究的目的；也可能查询的代价变大，这是我们所要避免的）。我们考察约束加入查询后我们获得的收益与付出的代价的总体情况。首先，看一下最极端的情况，即在约束引入后查询的约束条件产生矛盾，此时无须对数据库进行访问我们便知道查询的结果为空，或者查询的结果从约束条件直接得到，这样也无须访问数据库，在这种情况下收益是100%；如果在约束引入后，不能产生任何优化效果或反而增加了查询代价，则收益为0。

例如，对永久约束与状态约束而言，由于状态约束与数据库的状态密切相关，其优化效果要比永久约束好。其直观的原因在于永久约束是在任何数据库状态下都成立的完整性约束，因此，此类约束必然大多是很显而易见的领域规则。例如，所有的人的年龄都小于200岁，类似这样的语义规则，尽管正

确，但对于查询不会产生任何优化效果；状态约束则恰恰相反，其具有较强的时态性，所以大多数的约束都是用户发出查询时所不清楚的，因此，在进行语义查询优化时会得到较好的效果（此处，我们并没有考虑时态约束在数据库的状态发生变化时，所需要的维护代价，所以我们说时态约束较永久约束有较好的优化效果）。

然而，系统的定量的研究语义规则的内容和表现形式对语义查询优化的影响的工作目前还没有开展。只有文[17]对数据库中的完整性约束分类及特点，在语义查询优化的背景下进行了一定的深入的探讨。尽管如此，我们深信，这方面工作的成果会对语义查询优化的其他问题的研究起到促进，是一项很有意义的工作。

3. 语义规则的自动获取

研究者在语义规则的自动获取上做了大量的工作，在方法上主要可以分为两种：查询驱动和数据驱动的。在查询驱动的实现框架中[4~7]，语义规则的获取是通过语义等价查询的比较进行的（如果两个查询在同一个数据库状态下所得到的结果集是相同的，则说这两个查询是语义等价的）。通过这两个查询的约束条件的比较，便可以得到一些候选的语义规则集，再通过一些裁减技术，便得到了我们所需要的语义规则。查询驱动方法的最大的缺点是只有查询被重复时，以前获取的语义规则可能才是有用的；换言之，如果一个查询以前没有遇到过，则必须在线重新学习新的语义规则，将会产生很大的系统代价。

数据驱动[8~10]的方法主要是通过对数据库的数据进行分析，来得到大量的语义规则，这种方法可以借鉴人工智能、机器学习以及数据挖掘中的一些思想和方法。它的主要缺点是学习缺乏针对性，即学习到的语义规则可能对查询优化毫无用处。

在以往的研究工作中，有几种典型的方法分别在文[6，8，12]等中被提出来。第一种方法是 Siegel 使用的基于查询的方法，在他的系统中，一些非常简单的规则可以被学习出来。同时，他通过使用预先定义的启发式规则与实例查询来引导语义规则的自动获取。这些启发式规则与 King 在其博士论文[11]中用来引导语义查询优化过程的启发式规则是相同的。

例如：如果一个在属性 A 上的选择条件被另一个在属性 B 上的选择条件所蕴含，且 A 不是索引属性，则其在 A 上的选择条件可以从查询中删除。

这条启发式规则叫做选择缩减，还有另外7条其他的启发式规则。根据这条规则，当系统收到一条查询时，检查是否有选择条件其相关属性 A 非索引属性；如果 A 存在，则系统会试图学习行如启发式规则形式的语义规则。在这种情况下，假设系统收到一条带有 n 个选择条件的查询，且他们都不是在索引属性上，则系统仅仅根据这一条启发式规则就会试图验证 n^2 条语义规则的成立与否。此外，这种方法没有考虑到学习到的规则的鲁棒性。这种方法最根本的局限性是启发式规则仅仅依赖于特定的查询而没有考虑到数据库中数据的性质，因此可能会失去很多高可用性的语义规则。

Shekhar 等提出了一种数据驱动的方法来解决语义查询优化中语义规则的自动获取的问题[8]。他们的系统是基于一个这样的假设：语义规则可以从数据库中属性值的非均匀分布中得到。为了发现非均匀的数据分布，他们的系统通过构造数据分布窗格的集合来实现。如表1所示，就是一个数据分布窗格的例子。

表1　一个数据分布窗格的例子

BusinessType	CargoType: NaturalGas	CargoType: RefineOil	CargoType: Others
Petroleum	1230	300	0
NoPetroleum	356	30	523

数据分布窗格中的数值代表满足限制条件的元组数目。例如，满足 BusinessType = Petroleum 且 CargoType = RefineOil 的元组的数目是300。从这个窗格中，我们不难得到下面的两条语义规则：

(BusinessType = Petroleum) ⇒ (CargoType ∈ {NaturalGas, RefineOil})

(BusinessType = NonPetroleum) ⇒ (CargoType ∈ {NaturalGas, RefineOil, Others})

显而易见，对于一个给定的数据库，用于产生语义规则的潜在的数据分布窗格的数目是十分巨大的，必须采取一定的措施加以限制。文[8]通过人为地加入限制条件的方法来限制窗格的数目，这就使该方法产生了很大的局限性。

文[12]中提出了一种基于线形回归统计模型的语义规则自动获取方法，该方法简单且十分实用，已经在 IBM DB2的原型系统中实现。它首先是基于这样的一个前提，即在语义查询优化中能够产生重要作用的语义规则是那些结构简单且易于理解的。因此，他们限定只提取那些简单的在条件和结论中分别只包含一个属性的语义规则。对于任意的两个属性，如果其值域都是有序的，二者之间的关联关系可以用如下的线形关系来表示：

$Y = bX + a + [emin, emax]$

上面的表达式表示 $Y > bX + (a + emin)$ 且 $Y < bX(a + emax)$，其中 a、b 是常数，X、Y 是属性值，emin、emax 是用来定义范围值的。

不难看出，数据驱动和查询驱动的方法都有各自的不足之处：数据驱动的方法在进行语义规则获取时缺乏针对性，可能得到大量的无用的语义规则；查询驱动的方法获得的规则可能针对性太强而缺乏通用性，降低系统的效率。因此，能否提出一种新的方法，弥补两种方法的不足而具有二者的优点，是将来需要进一步研究的问题。同时，能够对语义查询优化起作用的语义规则具有何种特征？自动获取的规则集的大小是否应该有一个合适的边界界线？都是未被研究的问题。如果能够取得一定的进展，克服这个语义查询优化技术应用的瓶颈，将是一件十分有意义的工作。

4. 语义规则的维护

对于大量的时态约束而言，语义规则的维护是一个十分重要的问题。虽然不同的语义规则具有不同的鲁棒性，但在数据库的状态发生变化时它们的维护是不可避免的。当它们与数据库的状态不一致时，最简单的办法就是将其删除掉。但是这种简单的处理方式可能引起语义规则集的振荡，使系统付出更大的维护代价。

文[5]定义了语义规则的鲁棒性，使系统充分学习鲁棒性较高的规则，这样在语义规则与数据库的状态不一致时，只是简单地将它们删除，并不会引起语义规则集的振荡。但是，在规则的获取时，对规则鲁棒性的估计所付出的代价以及高鲁

棒性的语义规则的可用性都是需要考虑的问题。

可以考虑对不一致的语义规则进行改造，使其在改造后与数据库的状态相一致，这需要复杂的规则维护机制[4]。

5. 查询优化

这其中包括可应用规则的选取、等价优化查询的选取以及优化与执行的代价模型等问题，这是语义查询优化技术的核心问题。

语义查询优化的过程主要分为两个阶段：首先，根据原始的查询和语义规则集，选取可应用的语义规则并构造与原查询语义等价的查询的集合；然后，从中选出较初始的具有较小查询代价的查询来执行。显然，将与原始查询语义等价的所有的查询全部构造出来，然后从中选取出来一个最优的，代价十分巨大并且也是不现实的。因此，必须设计有效的算法来快速高效地完成查询优化任务。

最常用的技术就是对参加查询变换的匹配语义规则进行限制，通常是使用一些启发式规则，来减少产生的语义等价的查询的个数，其中包括：

·谓词引入：引入该谓词后可以使变换后的查询具有最小的执行代价。例如，在查询的索引属性或连接属性上引入谓词后，可以加快查询的执行速度。

·谓词删除：如果我们知道查询中的一个谓词在当前的数据库状态下，总是满足的，且执行改谓词代价较大，则我们可以从查询中将其删除。

·连接删除：如果一个查询包含的一个连接的结果可以通过语义规则获得，则我们就可以将该连接删除。例如，涉及到具有外关键字关系的两个表的连接就是此类情况。

下面介绍一些典型的方法。

King 在其博士论文[11]中给出7条启发式规则来进行查询优化，我们只列出其中的3条：

(1)索引引入规则 H1：如果查询中的关系 R 有一个没有约束的聚簇索引属性 B，则寻找那种隐含 B 上的约束的规则。

(2)减小搜索范围规则 H2：如果关系 R 上的唯一限制是一个联接条件，则寻找此关系上的选择条件。新的选择条件应该从所联接关系的选择条件中导出。

(3)选择缩减规则 H3：如果属性 A 上的选择条件被属性 B 上的另一选择条件隐含，而且 A 不是索引属性，也不是根据减小搜索范围规则引入的选择条件，则 A 上的选择条件可以从此查询中删除。

大部分目前的研究都是基于文[11]的启发式规则的方法的，在其基础之上进行了改进和提高[5,7,14,15]。值的注意的是，文[12]提出的方法很有新意，它给出了语义规则的"search ratio"的概念。每一条语义规则的"search ratio"由它的前提和结论的相关联的元组数目，即需访问的磁盘块数来决定的，这在语义规则的自动获取时是很容易得到的。在进行查询变换时，根据匹配的语义规则的"search ratio"来决定规则的应用与否。这种方法将严格地量化查询变换的规则，要比基于启发式规则的方法好得多。因此，很有应用价值和理论意义。

此外，一个经常被忽略的问题是优化与执行的代价模型问题。因为，与传统的代数优化相比，语义查询优化的过程相对较长，必须考虑其优化时间和执行时间的总和，来计算语义查询优化方法的收益。文[16]系统地研究了这个问题，并给出了一系列的解决方法。

结论 语义查询优化技术在某些情况下，可以显著地提高查询速度，这已经在某些商业的数据库产品中得到了应用[15]。因此，它是一个值得深入研究的领域。同时，我们必须认识到，只有有效地解决了上面所提到的问题，语义查询优化技术才会得到深入而广泛的应用。

参 考 文 献

1 Elmasri R, Wiederhold G. Data model integration using the structral model. In: Proc. Int. Conf. Management Data. ACM SIGMOD. 1979

2 Hammer M, Mcleod D. Semantic Integrity in a relation database system. VLDB, 1975

3 Codd E. Extending the database relational model to capture more meaning. TODS, 1979, 4(4)

4 Yu C T, Sun W. Automatic knowledge acquisition and maintenance for semantic query optimization. IEEE Trans. Knowledge and Data Engineering, 1989, I(3): 362～375

5 Hsu C N. Learning Effective and Robust Knowledge for Semantic Query Optimization: [PhD thesis]. Department of Computer Science, University of Southern California, 1996

6 Seigel M. Automatic rule deviation for semantic query optimization: [PhD thesis]. Boston University, 1988

7 Siegel M D, Sciore E, Salveter S. A method for automatic rule deviation to support semantic query optimization. ACM Transactions on Database Systems, 1992, 17(4): 563～600

8 Shekhar S, Hamidzadeh B, Kohli A. Learning Transformation Rules for Semantic Query Optimization: A Data-driven Approach. IEEE Trans. Knowledge and Data Engineering, 1993. 946～964

9 Han J, Cai Y, Cercone N. Data-driven discovery of quantitative rules in relational databases. IEEE Trans. Knowledge and Data Engineering, 1993, 5(1): 29～40

10 Sayli A, Lowden B G T. The use of statistics in semantic query optimization. The 13 European Meeting on Cybernetics and Systems Research, Vienna, April 1996. 991～996

11 King J. Query optimization by semantic reasoning: [PhD thesis]. Stanford University, Department of Computer Science, 1981

12 Gryz J, et al. Discovery and application of check constraints in DB2 IEEE International Conference on Data Engineering, 2001

13 Sayli A, Lowden B G T. A fast transformation method for semantic query optimization. IDEAS, 1997. 319～326

14 Shenoy S T, Ozsoyoglu Z M. Design and implementation of a semantic query optimizer. IEEE Trans. Knowledge and Data Engineering, 1989, I(3): 344～361

15 Cheng Q, et al. Implementation of semantic query optimization techniques in DB2 universal database. In: Proc. of VLDB, 1999. 687～698

16 Srivastava J, Dutta S. A formal model of trade-off between optimization and execution costs in semantic query optimization. VLDB 1988. 457～467

17 Godfrey P, et al. Exploiting constraint-like data characterizations in query optimization. SIGMOD, 2000