文章编号:1001 - 9081(2010)12 - 3371 - 03

基于社区划分的联机分析处理查询优化方案

何昭青[1,2],周 攀[2],杨科华[2]
(1. 湖南第一师范学院 信息科学与工程系,长沙 410205; 2. 湖南大学 计算机与通信学院,长沙 410082)
(hezhaoqing@126.com)

摘 要:针对 P2P 环境下的联机分析处理(OLAP)查询节点数目不断增加时,易造成网络拥塞、查询效率降低的问题,提出一种基于社区划分的 OLAP 查询优化方案。该方案构建一个虚拟的社区网,并在此结构上设计了一种基于社区划分的多维数据集(CPDS)的 OLAP 查询优化算法。实验结果表明,该算法能有效避免因网络节点数目递增而导致的网络负载加剧问题,能有效地减少网络拥塞,优化了 OLAP 的查询效率,进一步提高 P2P 环境下 OLAP 的决策分析性能。

关键词:联机分析处理;数据立方体;点对点网络;虚拟社区
中图分类号: TP311 **文献标志码**:A

Community-partition-based online analytical processing query optimization

HE Zhao-qing[1,2], ZHOU Pan[2], YANG Ke-hua[2]
(1. *Department of Information Science and Engineering, Hunan First Normal College, Changsha Hunan* 410205, *China*;
2. *College of Computer and Communication, Hunan University, Changsha Hunan* 410082, *China*)

Abstract: In the Peer-to-Peer (P2P) environment, when the number of nodes of On-Line Analysis Processing (OLAP) query increase, network congestion will be aggravated and OLAP query efficiency will be reduced. Therefore, this paper proposed an optimized OLAP query method based on community partition. A visual community network was constructed with the method, and an algorithm of Community Partition Data-cube Search (CPDS) was designed in this structure. The results of experiment show that this algorithm can effectively avoid increasing network burden, when network OLAP nodes increase. Therefore, this method reduces congestion of network and optimizes efficiency of OLAP query, which improves the performance of decision-analysis of OLAP in P2P environment.

Key words: On-Line Analysis Processing (OLAP); data cube; P2P network; visual community

0 引言

由于 P2P 网络具有快速访问、避免单点失效和中心节点瓶颈等优良特性,因此被广泛应用于网络节点的计算和资源存储。近几年来,特别在数据挖掘的数据仓库技术中,利用基于 P2P 技术进行联机分析处理(On-Line Analysis Processing, OLAP)成为研究热点问题。对于 B/S 模式下 OLAP 服务器负载不断加剧导致 OLAP 查询效率低下的问题,有学者提出采用动态缓存的方法来解决[1-2]。然而,B/S 模式下对数据的获取是存在性能瓶颈的,因此有人提出采用 P2P 技术来提高 OLAP 的决策分析效率[3],并在对 P2P 模式下的多维数据集的维层次进行严格限制的条件下提出了一种数据立方体的模式匹配方案[4]。多维数据集维层次的限制给多维数据集的查询效率带来了一定的制约,于是文献[5]介绍了 P2P 网络环境下 Data Cube 模式匹配算法,在 P2P 环境下的多个 OLAP 服务器共同协作的前提下优化了 OLAP 的联合查询。但是在 P2P 环境下采用洪泛式搜索方式进行 OLAP 决策分析[6-7],当节点数目不断增加时极易导致网络拥塞,使得 OLAP 决策效率大打折扣。因此,本文通过对 OLAP 网络节点进行社区划分后,构建了一个虚拟社区网络,提出了一种基于社区划分的多维数据集(Community Partition Data-cube Search, CPDS)算法,进一步提高了 OLAP 的查询效率。

1 虚拟社区网的构建

在 OLAP 网络中实现多维数据集的查询分析,是通过将用户节点的查询需求发送到 OLAP 网络中,再由邻居节点进行转发搜索来完成的。文献[7]中的 DQDC 算法就是根据 P2P 环境下的洪泛式搜索方式来进行设计的。

因此,当 OLAP 网络中的用户不断增加、规模不断扩大的时候,利用洪泛式搜索来进行多维数据集的定位查询将会带来很大的网络开销。这种搜索机制虽然能够将查询包完整地覆盖到整个 OLAP 网络,但是当网络中存在环路的时候就会明显降低整个系统的决策分析效率。本文将整个 OLAP 网络中的节点按照其查询的特性进行划分虚拟社区,通过虚拟社区网来提高整个系统的查询分析性能。

1.1 虚拟社区网

节点的查询特性也就是该节点在一段时间内对同一多维数据集的查询频率,本文将此定义为该节点的特征值。因此,根据 OLAP 网络中节点的一系列查询条件,就可以将整个 OLAP 网络节点划分为许多特征值不同的虚拟网络区域。据此,本文提出将整个 OLAP 网络划分为两个逻辑上独立的虚拟网:一是基于特征值发布的社区网;二是基于节点提供多维数

收稿日期:2010 - 05 - 24;**修回日期**:2010 - 07 - 25。
基金项目:广东省产学研项目(2007A090302079);湖南省教育厅科研项目(08C016)。
作者简介:何昭青(1964 -),女,湖南邵阳人,教授,主要研究方向:计算机优化算法、Web 3D、网络教育; 周攀(1982 -),男,湖南岳阳人,硕士研究生,主要研究方向:数据库、数据仓库; 杨科华(1979 -),男,湖南邵阳人,副教授,博士,主要研究方向:数据库、嵌入式软件。

据集查询分析的资源网。当然两个网络的节点能覆盖整个 OLAP 网络。本文将虚拟社区网中特征值相同的节点集定义为社区。那么一个社区当中的所有节点拥有相同或者相似的特征值。图 1 为 OLAP 网络按照社区划分的虚拟网络结构。

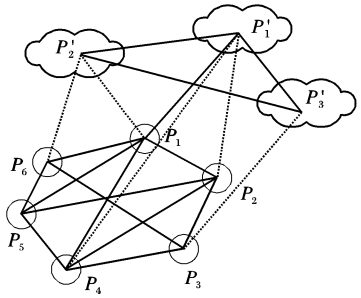

图 1 基于社区划分的 OLAP 虚拟网络结构

从图 1 可以看出，由 $\{P_1,P_2,P_3,P_4,P_5,P_6\}$ 六个节点构成的网络叫做资源网；P_1,P_2,P_4 三个节点构成社区 P_1'，P_1，P_6 二个节点构成社区 P_2'，P_3 单节点构成社区 P_3'，而由 $\{P_1',P_2',P_3'\}$ 三个社区中的节点的并集 $\{P_1,P_2,P_3,P_4,P_5,P_6\}$ 构成的网络叫做虚拟社区网。对 OLAP 服务器节点，由于其功能只单一为其他节点提供服务，所以单独作为一个社区进行划分。

1.2 网络节点的特征值

在实际的应用当中，节点的查询需求可能是需要对一个多维数据集进行一系列的聚集查询分析后才能得到查询结果。同时由于网络中的每一个节点的查询需求是多变的，可能对于同一多维数据集，在同一维的不同层次上进行聚集都可能得到不同的结果，当然就不能按照查询需求构建的语义维层次链[5]或者扩展语义维层次链[7]来进行社区的划分。因此，本文将多维数据集作为 OLAP 网络中节点的特征值表示。表 1 为一个 OLAP 网络中节点社区划分的特征值表示。其中多维数据集的表示是通过一个随机生成的唯一二进制串来完成的。

表 1 OLAP 节点特征值表示

节点 ID	多维数据集 ID	搜索%	社区 ID
00（P_1）	1011(空调销售)	12	0110（P'_1）
01（P_2）	1011(空调销售)	24	0110（P'_1）
11（P_6）	1010（冰箱销售）	18	0111（P'_2）
…	…	…	…

对表 1 中的数据进行分析，可以看到同一个节点可能同时处在几个不同的社区当中。从另一个角度说明节点的查询需求是多方位的。事实上，一个节点的查询需要分析的多维数据集可能有很多，本文提出将节点对同一多维数据集的查询频率和更新时间作为该节点特征的衡量标准。每一个节点，需要按照自己的特点对多维数据集的查询频率和更新时间进行排序，然后再取排序在前面的多维数据集（具体个数根据多维数据集自身的特性来决定）作为自己的特征值。随着节点查询需求的变化，节点的特征值也就随之发生变化。

1.3 基于特征值的虚拟社区网构建

基于上文对整个 OLAP 网络的划分进行详细的阐述后，根据 P2P 环境的高度动态性，如何实现每一个 OLAP 节点加入和退出整个虚拟社区网是实现 OLAP 查询的关键。

用户的需求是变化的，多维数据集的数据随着时间的积累也是不断更新的，那么社区网中的特征值当然也就需要随之而发生变化。因此对于社区网中每一个社区的节点管理是实现整个多维数据搜索的关键。本文通过在每一个社区中建立一个共享式的路由链表[8]实现对整个社区节点的特征值动态变化的管理。表 2 为社区共享式路由链表。

表 2 社区节点共享式路由链表

社区 ID	节点 ID	节点 IP	特征值
0110（P'_1）	01（P_1）	127.0.0.1	1011(空调销售)
1000（P'_3）	10（P_5）	127.0.0.1	1000（彩电销售）
…	…	…	…

在每个社区中每个节点处都会存储一张类似的链表，并将特征值建立为该链表上的索引。根据 OLAP 网络中节点高度动态性的特点，当社区中的一个节点离开社区时，就只需要更新其他节点的路由链表。社区的建立是为了避免诸如洪泛式搜索方式的网络冗余，提高 OLAP 决策分析的性能。下面结合一个例子来说明一个已经加入 OLAP 网络节点发布特征值的过程。

例 1 根据表 1 和表 2 的数据以及图 1 的结构进行介绍，若节点 P_3 对多维数据集 1011 的查询频率由原来的 24 增加到 46，而对 1000 的查询频率由原来的 18 降到 15。且对 1011 更新的时间为 20100301，此时节点 P_3 发送一个申请社区注册信息包。当该信息包在社区中路由至社区 0110 时，该社区根据特征值的信息更新社区 0110 的路由链表信息。同时节点 P_3 还发送一个注销社区信息包到节点退出的社区网，该社区再根据信息包的内容更新路由链表信息。通过如图 2 所描绘的特征值发布算法结构流程，形象化地说明了特征值发布的过程。

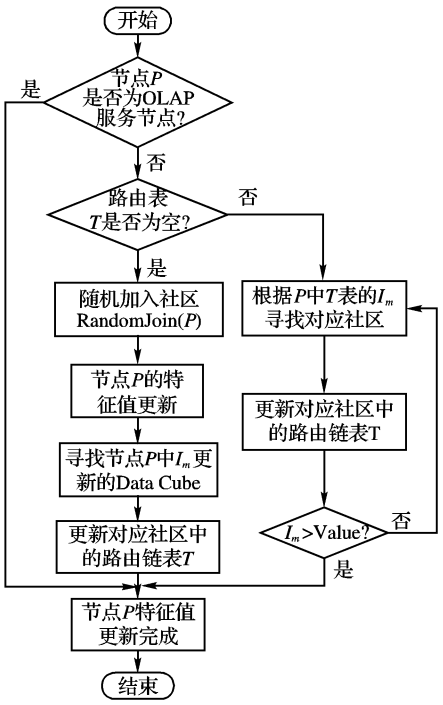

图 2 特征发布算法结构流程

该算法的设计是基于特征值发布的模拟[6]。假设节点 P 拥有的路由链表为 T，P 向社区发布的特征值集合为 $I_m=\{D_1,D_2,\cdots,D_m\}$，$m$ 的大小由每个发布节点决定。节点 P 向社区发布特征值（Node Feature Release，NFR）算法的伪码设计如下：

1）根据新加入的节点 P 和该节点的特征值，进行社区划分；

2）当搜索到的 *Object* 集合不是服务 OLAP 服务节点,且节点 P 的特征值已知时,判断 P 的特征值 I_i 和路由表中的节点的特征值 I_i' 是否匹配;

3）如果匹配($Q \geqslant Value$),将该节点的特征值赋值给该节点的阈值 *Value*;如果不匹配,就搜索下一个节点;

4）根据搜索到的 *Object* 集合,查询该集合中节点的特征值的更新时间;

5）如果更新时间在规定的时间范围内,且 P 不在 *Object* 集合中,就将节点 P 加入到 *Object* 集合中 $Join(P, Object_i)$;

6）如果节点 P 的特征值未知,则将节点 P 随机加入到一个社区中 $RandomJoin(P)$;然后递归调用 NFR 算法;

节点的特征值发布完成以后,如果节点的特征值发生改变,该节点 P 可以通过 NFR 算法向社区重新发布特征值。这里需要注意的是作为 OLAP 服务器的节点,在加入到网络后都是一种稳定的状态,因为它不需要加入和退出网络。

2　基于社区划分的 CPDS 算法设计

根据上文的分析,OLAP 网络分逻辑上两个独立资源网和虚拟社区网,OLAP 的决策分析都是需要整个 OLAP 网络节点的协同工作才能完成。如何有效使用这两个网络协同高效率工作,本文设计了 CPDS 查询分析算法。

一个节点怎样在社区中完成它对多维数据集的决策分析呢? 首先,节点 P 对用户进行的决策分析语句 Q 进行语义和扩展语义上[7]的多维分析;根据多维分析的结果,从中获取 Q 分析的多维数据集和该多维数据集对应的社区 ID;然后根据社区 ID 从节点 P 的路由表中查找对应得社区 ID,若找到对应的社区 ID,则根据此社区 ID 从社区虚拟网中进行搜索,最后从找到的社区中将该区域中的节点 ID 加入到节点 P 的路由表中并从这些节点处获取决策分析语句 Q 所需要的数据;若没有找到对应的社区 ID,则需要将对应多维数据集的 ID 根据 NFR 算法发布到社区虚拟网中,再从社区虚拟网中搜索相应的社区并从该社区的节点中下载用户需要的数据。这就是节点 P 在社区划分的前提下进行多维数据分析的整个过程。为了详细了解节点 P 进行数据分析的过程,本文从算法的执行流程来了解,其结构流程如图 3 所示。

从图 3 可以看出,该算法实现多维数据分析的关键就是对社区的查找,和节点对自己特征值的更新。该算法的伪码设计如下:

```
void CPDS( Q )//算法的伪码描述, Q 为节点 P 发起的查询分析语句;
{
 分析 Q 中的语义层次链和扩展语义层次链;
 Get( Data Cube ID, 社区 ID);
 Ret = Query(社区 ID, 节点 P);      //从路由表中寻找社区 ID
 if ( Ret )
   {从该社区的节点处获取数据,再进行数据的决策分析};
 else
 {
  从虚拟社区网中寻找对应的社区;
  NFR( P, Value );                  //发布节点特征值
  从该社区中的节点处获取数据,再进行数据的决策分析;
 }
}
```

总之,CPDS 算法优化了 OLAP 网络中洪泛搜索算法。CPDS 算法是将多维数据集的查询方式由原来的洪泛式搜索改变为先从虚拟社区网中获取对应的社区 ID,再从社区的节点集中获取用户需要数据。这样,减少了 OLAP 的网络冗余,提高了决策分析的效率。

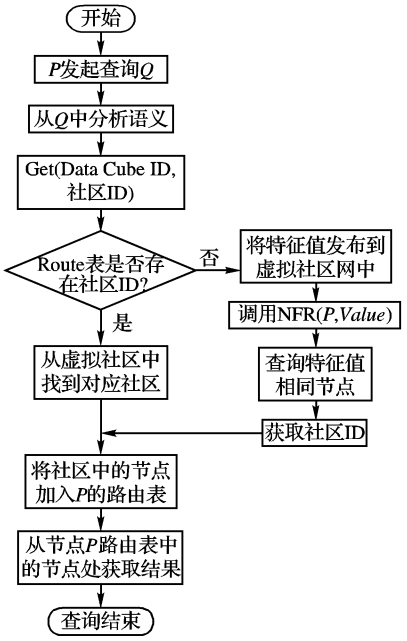

图 3　CPDS 算法结构流程

3　实验结果和性能分析

为了测试 CPDS 算法的性能,下面从实验的角度来证明该算法不仅可以高效地实现 OLAP 网络环境下的多维数据分析,而且还能避免随着用户的增加,网络的负载也随之增加的弊端。并对 CPDS 算法和文献[7]的 DQDC 算法以及传统的 OLAP 查询算法在 OLAP 的决策分析效率上进行对比,结果如图 4 所示。

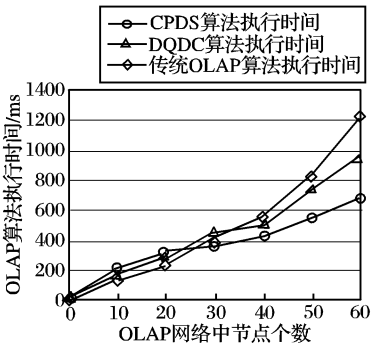

图 4　CPDS 算法和 DQDC 算法性能比较

在进行测试之前,搭建了一个 P2P 网络,网络中搭建了一个 Web 服务器,一个 Tracker Server 服务器,其余节点为 OLAP 网络节点(为简单起见,每个 OLAP 网络节点上只有两个数据立方体)。同时,从 OLAP 网络节点中构造一个虚拟的社区网。

从图中可以看出,CPDS 算法与其他 OLAP 查询算法在节点数比较少的时候,其 OLAP 的查询分析时间基本都差不多,有时候 CPDS 算法的执行时间可能还比较长。因为 CPDS 算法在社区虚拟网中进行多维数据集的查询和社区中每一个节点处进行路由链表中节点数据更新等需要消耗部分的时间。然而,当节点的数目不断增加的时候,CPDS 算法比其他算法的性能更好。这样不仅提高了 OLAP 决策分析的执行效率,而且更加完整地实现了 OLAP 网络环境下的多维数据分析,避免了 OLAP 网络的冗余。

(下转第 3376 页)

$$D_A(x,y) = \begin{cases} \{a_i \in A : (x,y) \notin G(a_i)\} : (x,y) \notin G(A) \\ \phi : \text{其他} \end{cases}$$

因此，$D_A(x,y)$ 是个体 x 与 y 在关系 G 下有区别的所有属性的集合。

定义8 令 GI 为不完备灰色信息系统，定义 $\Delta = \bigwedge_{D_A(x,y) \in D} \vee D_A(x,y)$ 为 GI 中区分函数。

利用布尔推理技术，可将其化为极小析取范式。在其极小析取范式中，每个合取子式就对应属性集合 A 的一个约简，所有合取子式就是 A 的全部约简。全部约简的交即为 A 的核。

根据所介绍的属性约简的方法，取常量 $\varepsilon = 0.5$，表2所对应的区分辨识矩阵为表3。通过定义，GI 的区分函数为 $\Delta = a_1 \wedge a_2 \wedge a_4$，则 $B = \{a_1, a_2, a_4\}$ 是表2的一个约简。

表3 关于表2的区分辨识矩阵

$x_i \backslash x_j$	x_1	x_2	x_3	x_4	x_5	x_6
x_1	$\varnothing$	a_1	$\varnothing$	$\varnothing$	a_1	a_1
x_2	a_1a_2	$\varnothing$	a_1a_3	a_4	$\varnothing$	$\varnothing$
x_3	a_4	a_1a_3	$\varnothing$	$\varnothing$	a_1	a_1a_3
x_4	a_4	a_1a_4	a_4	$\varnothing$	$\varnothing$	a_4
x_5	a_1a_2	$\varnothing$	a_2	$\varnothing$	$\varnothing$	$\varnothing$
x_6	a_1a_2	$\varnothing$	$a_1a_2a_3$	a_1	$\varnothing$	$\varnothing$

4 结语

由于现实生活中存在着信息和知识的不完全性和复杂性，促进了各种拓展粗集模型的产生与发展。本文以部分信息已知、部分信息未知的小样本、贫信息、不确定的系统为研究对象，定义了不完备灰色信息系统，据此提出了变精度灰相似关系及相应的粗集模型，并给出了一些基本性质以及计算约简的实际操作方法，因此本文的工作对拓展区间值信息系统的粗集模型有着重要的意义。

参考文献：

[1] PAWLAK Z. Rough sets theory and its applications to data analysis [EB/OL]. [2010-05-01]. http://www.cs.uakron.edu/~chan/cs460/Spring%202005/Rough%20Set%20and%20Its%20Applications.pdf.

[2] PAWLAK Z. Rough sets and intelligent data analysis[J]. Information Sciences—Informatics and Computer Science: An International Journal, 2002, 147(1/4): 1-12.

[3] 王国胤. Rough集理论在不完备信息系统中的扩充[J]. 计算机研究与发展, 2002, 39(10): 1238-1243.

[4] LEUNG Y, LI D Y. Maximal consistent block technique for rule acquisition in incomplete information systems[J]. Information Sciences, 2003, 153(1): 85-106.

[5] LEUNG Y, WU W Z, ZHANG W X. Knowledge acquisition in incomplete information systems: a rough approach[J]. European Journal of Operational Research, 2006, 168(1): 164-180.

[6] QIAN Y H, LIANG J Y. Positive approximation and rule extracting in incomplete information systems[J]. International Journal of Computer Science and Knowledge Engineering, 2008, 2(1): 51-63.

[7] KRYSCKIEWICZ M. Rough set approach to incomplete information system[J]. Information Sciences, 1998, 112(1/4): 39-49.

[8] STEFANOWSKI J, TSOUKIAS A. Incomplete information tables and rough classification[J]. Computational Intelligence, 2001, 17(3): 545-566.

[9] GRZYMALA-BUSSE J W. Data with missing attribute values: Generalization of indiscernibility relation and rule induction[C]// Transactions on Rough Sets, LNCS 3100. Berlin: Springer-Verlag, 2004: 78-95.

[10] GRZYMALA-BUSSE J W. Characteristic relations for incomplete data: a generalization of the indiscernibility relation[C] // Rough Sets and Current Trends in Computing, LNCS 3066. Berlin: Springer-Verlag, 2004: 244-253.

[11] DENG J L. Control problems of grey system[J]. System & Control Letter, 1982, 1(5): 288-294.

[12] 刘思峰, 郭天榜, 党耀国. 灰色系统理论及其应用[M]. 3版. 北京: 科学出版社, 2004.

[13] YANG XIBEI, YU DONGJUN, YANG JINGYU, *et al*. Dominance-based rough set approach to incomplete interval-valued information system[J]. Data & Knowledge Engineering, 2009, 68(11): 1331-1347.

[14] QIAN YUHUA, LIANG JIYE, DANG CHUANGYIN. Interval ordered information systems[J]. Computers & Mathematics with Applications, 2008, 56(8): 1994-2009.

[15] WU S X, HUANG Z Y, LUO D L, *et al*. A grey rough set model based on (α, β)-grey similarity relation[C]// Proceedings of IEEE International Conference on Grey Systems and Intelligent Services. Nanjing, China: IEEE Press, 2007: 903-909.

(上接第3373页)

4 结语

本文主要是从P2P网络的高度动态性这个角度出发，探讨了OLAP网络模式下多维数据分析的决策效率。通过构建虚拟社区网络模型，实现了基于社区划分的OLAP查询，提高了OLAP的决策分析效率，极大地降低了OLAP网络的负载。在以后的相关研究中，将在数据立方体的共享单元和更新等关键问题上进行深入的研究和探讨。

参考文献：

[1] 曹丽娟, 谢强, 丁秋林. 基于分布式数据缓存技术Web-OLAP系统研究[J]. 计算机应用, 2008, 28(2): 515-518.

[2] KALNIS P, WEE S N, BENG C O, *et al*. An adaptive peer-to-peer network for distributed caching of OLAP results[C]// Proceedings of the ACM SIGMOD Conference. New York: ACM Press, 2002: 25-36.

[3] TATARINOV I, HALEVY A. Efficient query reformulation in peerdata management systems[C]// Proceedings of the ACM SIGMOD International Conference. Paris, France: ACM Press, 2004: 539-550.

[4] ESPIL M M, VAISMAN A A. Aggregate queries in peer-to-peer OLAP[C]// Proceedings of Seventh ACM International Workshop Data Warehousing and On-Line Analytical Processing. Washington, DC: ACM Press, 2004: 102-111.

[5] 杨科华, 魏莉. 一种P2P网络环境下的OLAP模式匹配方案[J]. 计算机工程与应用, 2008, 44(9): 162-164.

[6] UPADRASHTA Y, VASSILEVA J, GRASSMANN W. Social networks in peer-to-peer systems[C]// Proceedings of the 38th Annual Hawaii International Conference. Kona, Hawaii: [s.n.], 2005: 200-211.

[7] 周攀, 杨科华, 周利民. 一种P2P网络环境下的OLAP查询方案[J]. 计算机工程与应用, 2010, 46(33): 140-144.

[8] TAN Y H, CHEN Z P, LIN Y P. Research and implementation on searching mechanism based on interest mining in unstructured P2P systems[J]. Computer Applications, 2006, 26(5): 1164-1166.