全球化信息系统的数据模型和查询研究

周晓华　（首都经贸大学信息系　100026）

摘要：网络上的用户使用大量散布在世界各地的信息源，然而介面上的不一致性给有效地存取信息带来许多困难，为了给用户提供一个统一介面的信息空间，本文提出了面向全球化网络处理的信息系统的概念框架，讨论了全球化信息查询的数据模型、地点描述、信息源、动态查询生成等关键内容，并给出了有关查询的优化算法。

关键词：全球化信息系统　信息源　动态查询计划

一、引言

目前信息和网络技术的飞速发展使得全球化的信息系统的建立成为可能而且十分必要，全球化的信息系统与传统的信息系统主要的区别在于前者的信息源不仅数量庞大，而且散布在世界各地。本文提出了面向全球化网络处理的信息系统的概念（Global Information Systems）框架，类似这样的全球化系统能够支持在线的信息检索任务，为成千上万的用户提供自助终端式的服务。

1.信息源

全球化信息系统的目的是提供给用户一个通用的信息源，当用户提出查询问题后，系统能够决定从哪个信息源找到相关信息并把结果提供给用户。类似这样的系统与一般的信息系统相比，突出的问题是信息源的数量十分庞大，如不进行有关的进一步处理，则会大大降低查询的效率。此外还要考虑信息源的一些其他特性，包括：

·信息源自动化问题：信息源必须能够实现更新数据，如果它能够提供有关其内容以及内容介面的描述，那么信息源的数据可以不必改变内部操作以适应全球化信息系统的需要。

·信息源动态特性：一个可用的信息源必须能够经常发生变化，比如加入新的信息源，删除无用的信息源。

·存取费用：一般来说，在网络上存取信息源的费用较高，所以系统在设计上要尽量考虑快速应答。

此外，在用户介面设计，要考虑应用多媒体技术，用户介面的屏幕设计、存取路径必须能够吸引用户，能够提供快速存取和多项服务，并且随着市场的不断发展，系统要不断适应新的情况，以满足大量终端用户的需求。

2.全球化信息系统结构

我们提出一个全球化信息系统结构如图 1 所示。

该系统能够给用户在概念上提供一个统一的信息空间，我们把这个称为全球化视图。用目标导向的数据模型来表示。在系统中，用户能够使用一个陈述性查询形式阐述用户的查询地点和存取细节，为了对用户的提问形成有效应答，全球化信息系统的查询处理器首先要使用一些与地点有关的描述内容，这些描述可通过地点语言给出，它包括地点的语义关系、表示地点信息的完整内容描述能力以及说明一个信息资源的能力。为此，有效地减少查询中不必要的信息源，从而加速整个查找过程。

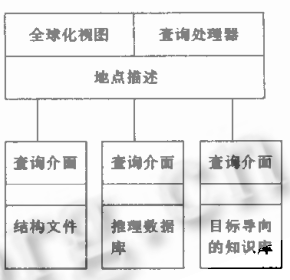

图 1

二、信息空间上的数据模型和查询形式

1.数据模型

我们的观点是用一个面向目标的数据模型来表示信息，它由 N 元关系、概念和限制这几类实体组成。

(1)N 元关系：这些关系的属性值由简单类型（整数、串）导出，或由复杂类型（定义为类的概念）导出，我们把 N 元关系记为 ϵ。

(2)概念（类概念）：描述了属于某个概念的一类客体所具有的共同特性，它可以通过层次结构来表示。用 D 代表类概念，则 $W = D \cup \epsilon$ 代表了我们研究的关系对象。

(3)限制：以全球化观点建立的数据模型的重要之处

在于它可以使用限制来表示丰富的语义知识，例如序列限制，α1θ α2，α1 和 α 2 代表变量或限制 θ∈(，〈，〈=，≠，=)，α1θα2(如 α1 为费用，α2 为＜1000，则 α1θα2 表示费用＜1000 这个限制)。

例 1：假定一个用户希望得到旅行社和航空公司在某一条飞行线路上最低机票信息，显然旅行社的服务电话可以提供这方面的有关信息，但用户可能需要一个个地去查找不同的数据库，而这需要花费许多时间、精力和金钱，有时信息不灵便无法达到目的。为了解决这个问题，我们建立价格数据库，通过抽取其中的有关事实来得到用户需要的信息，在该数据库中，允许旅行社使用不同的模式来表示它们的信息。

对于例 1 的情况，我们将有关概念形式化如下：

①报价关系：Quote(Ag, AL, Src, Dst, C, D)定义了一个旅行社 Ag 对于航空公司 AL 在日期 D 这天从 Src 到 Dst 的飞机票报价为 C。

②显示地区、电话号码关系：Dir(Cust、Ac、Tel-NO)列出了顾客 Cust 所居住地区代码 Ac 和电话号码 Tel-NO。

③显示旅行社名称关系：Name(Ag、Name)给出了旅行社 Ag 的名称 Name。

④代码关系：Areacode(Pl、Ac)定义了地点 PL 所在的地区代码 Ac。

限制和概念是用来说明全球化视图的关系的属性类型。例如显示关系“DIR ”中顾客属性 Cust 是通过顾客类型定义的。报价关系“Quote”的属性 Ag 是通过旅行社的类别定义的，如商业的子概念。其中报价关系的属性 C 值被限定为一个非负数。

在我们的这个方法中，使用类知识表示语言作为数据模型的一部分，通过一个可描述逻辑系统构造一个全球化信息系统，这是因为它具有较好地扩展能力，允许产生新的概念并可以自动地将新概念放到概念层次上，例如，假定在概念层中存有商业旅游和航空公司旅行社这两个概念，它们分别是商业类子概念，如果用户想加入一个新概念－－商业旅行社，那么类概念可以很轻易地把这个概念加到商业和航空公司旅行社之间的概念层上去。此外，支持类概念的系统，系统不要求用户明确说明与目标有关的所有概念，系统概括与目标有关的信息和概念并可以重新以适当的形式划分。

2.查询形式

在本文中，我们考虑一个表示形式的查询形式：

Q(X)：C(Y)，E1(Xi)，…，Ek(Xk)

其中，O(X)是查询命题名称，C(Y)是查询变量的序列限制合并，Ei 是全球化视图关系 W 的关系名称，X、Y、X1，…，Xk 定义了限制的元组、对象和变量。

例 2：假定要查询 T 城市 T1 地区的旅行社名称和电话，该旅行社提供从 A 国 A1 城市到 B 国 B1 城市的任何航空公司在 1000 美元以下的飞机票情况。

则以上查询信息可以表示成：

Query(Name、Ac、Tel-NO)：Areacode(‘T、T1’，Ac)，Quote(Ag，AL，‘A、A1’，‘B、B1’，C，D)，C＜1000，dir(Ag，Ac，Tel-NO)，Name(Ag，Name)

三、优化处理与查询估价算法

全球化信息系统的查询处理器采用的基本思想是：当用户提出查询问题后，系统首先针对查询进行基本分析，然后再决定从哪里可以得到信息并把结果提供给用户，在这个过程中，将用户查询分解成若干子查询，最后一组子查询的结合体形成对用户的最终应答。

由于在网络上存取费用较高，因此首先需要极小化与问题有关的外部地点关系，通过以下两种方法来达到这一点：在地点描述中使用限制方式排除那些与查询无关的内容和使用地点关系完整性信息删除冗余信息。

此外，在传统的数据库查询处理过程中，查询计划是在一定条件、时间背景下产生的，并不能随时修改，但是在一个全球化信息系统中，无法固定一个背景知识建立完整的查询计划，为此，我们提出一个算法，利用它可以使用实时信息在查询估价时删除后续的信息源。

1.地点和地点描述

虽然我们可以借助于 W 表示查询，但也应当看到，全球化视图关系仅仅构成信息空间概念表示，为了回答用户查询，为了删除非法信息源，系统应该具备地点关系的描述，我们认为一个地点描述由两类信息组成：

(1)内容：一个地点描述与从关系 W 扩展为 R 的内容的语义有关。

(2)能力：一个地点描述指出了基于关系 R 的查询种类。

例 3：一个旅游信息源通过扩展关系“travel－dir(Name，Ac，Tel-NO)”为旅行社提供信息服务。

从这个例子可以看到，该关系的内容描述为：包括了 dir 关系中的旅行社的电话号码信息；在能力方面描述了信息源能够回答两种查询的能力。第一，对于给定的旅行社，信息源能够提供了旅行社的名称、地区代码和电话。第二，信息源能够提供了所有旅行社的名称、地区代码和电话，而不能提供属性之外的其他信息。

2.信息源

我们通过提出一项计划完成将查询中的子目标联合起来,形成用户的有效应答,由于网络操作费用较高,因此要执行的主要优化手段是尽量减少回答用户查询的外部信息源的数量。

在这一节里讨论如何通过地点描述决定与查询有关的信息源。假定查询形式如下:

Q(X):C(Y),E1(Xi),…,Ek(Xk)

通过下面具体步骤决定相关的信息源:

对于每个 1≤i≤k,执行以下步骤,为了回答查询,通过对限制 Cq 在 X i 上的投影实现决定 Ei 有效成份(用 CEi 表示)。

例 4:假定外部关系 bus - 212 - dir 包含在 212 地区顾客的电话号码,同样 908 - dir 包括在 908 地区顾客的电话号码,有:

bus - 212 - dir(Cust, Tel-NO) ⋂ dir(Cust, 212, Tel-NO)

908 - dir(Cust, Tel-NO) ⋂ dir(Cust, 908, Tel-NO)

那么针对想找出 212 地区旅行社的电话号码这个查询就可以表示为:

Query(Tel-NO):dir(cust, 212, Tel-NO), TravelAgent(cust)

使用以上算法,地点关系 bus - 212 - dir 被认为与计算 dir(Cust, Ac, Tel-NO)有关,并满足 Ac = 212. and. TravelAgent(cust),而地点关系 908 - dir 就不被认为与这项查询有关,从而不被检索到。

3.**动态查询估价算法**

在全球化信息系统中,我们提出根据事实动态生成查询计划。

例 5:假定要检索 T 城市 T1 地区旅行社的电话号码,查询表示为:

Query(Ac, Tel-NO):Areacode("T、T1", Ac), TravelAgent(Ag), dir(Ag, AC, Tel-NO)

其中,TravelAgent(Ag)限制了在查询中包括与旅行社无关的信息源目录,也就是说,它只查找旅行社的名字但不查找其他信息。一般来说,如果没有这种上下文语义限制,在以全球化视图表示的知识元组中,查询计划可能把其他的目标信息认为与该查询有关。一旦子查询 Areacode("T、T1", Ac)被估价后,则查询严格限制在与问题有关的描述中。

例 6:假定 TM 和 TN 旅行社不是旅游旅行社的下属子概念,并假定有两个外部信息源:TM - Quotes(Ag, AL, src, Dst, C, D)和 TN - Quotes(Ag, AL, src, Dst, C, D),它们分别定义了各自的飞机报价。现要查找提供从 A 国 A1 城市到 A2 城市的机票价格 < = 500 的旅行社信息,相应的查询表示为:

Query(Ac, Tel-NO):Quote(Ag, AL, "A、A1", "A、A2", C, D), dir(Ag, Ac, Tel-NO), C< = 500

为了回答这个问题,首先在报价数据库 Quote 中查找旅行社的名字,然后在目录数据库中找到它们的电话号码,但是注意到在报价数据库中找到的旅行社可以有助于减少查找无关的旅行社,为此我们提出了一个称为动态查询估价的算法,它可以使用实时信息动态回答一个查询问题,基本思想是:只有在前面的子目标被估价后,后续子目标有关的信息源才动态生成。

动态查询估价算法(Q(X), SD):

其中 Q(X)是一个查询,其描述形式为(Q(X):C(Y), E1(Xi), …, Ek(Xk)), SD 是地点描述的集合。

步骤 1:决定 E1(X1), …, Ek(Xk)在查询 Q(x)中的顺序。Pi 定义为元组(t, Ct(y))的集合,其中 t 是关系 E1, …, Ei 中的一个元组,Ct(y)是限制。

步骤 2:从 i = 1, …, k 执行下面过程,对于 P 中的每个元组(t, Ct(y))∈ Pi - 1 do

①Ci(Xi)定义为 Ct(y)在变量 Ei(Xi)上的投影。根据地点描述 SD,计算关系 Ei(Xi)(满足限制 Ci(Xi))的元组 ti, t·ti 表示结合了第 i 个子目标的结果元组。

②C'i(Xi)定义为 CESD ∧ CRSD 在变量 Xi 上的投影,而 CESD 和 CRSD 是地点描述 SD 在两个方向的限制。如果一个元组(t·ti, C)已经在 Pi 中存在,则用元组(t·ti, C ∧ C'i(Xi))替换它,否则在 Pi 中加入元组(t·ti, Ct(y) ∧ C'i(Xi))

步骤 3:将 Pi 中的每个元组投影到 Q(X)上就可以得到对查询的回答。

参考文献

[1] Xu Wu and Nick Cercone A Knowledge - Based System for Generating Informative Responses to Indirect Database Queries, J. of Intelligent Information Systems 1995, 5, pp5 - 23.

[2] Alony. leyy, Diveshsrivastava, Thomas kirk" Data Model and Query Evaluation in Gobal Information Systems, Journal of Intelligent Information Systems, 1995, 5, pp121 - 143.

(来稿时间:1998 年 2 月)