

基于 CWM 商业报告多维分析模型的构建

朴 勇^{1,2} 王秀坤¹ 陈志奎² 王 征²

(大连理工大学电信学院 大连 116023)¹ (大连理工大学软件学院 大连 116620)²

摘 要 随着财务报告的规范化和正规化,大量企业开始使用扩展商业报告语言(XBRL)标准进行财务数据的披露与管理。但 XBRL 结构复杂并具领域相关性,缺乏进行数据分析的语义描述,从而导致以 XBRL 格式存储的数据无法被现有的多维数据分析工具所理解和利用。为此分析了 XBRL 的结构,应用了与平台及领域无关的通用仓库元模型(CWM)标准,给出了一种将 XBRL 格式的财务报告映射为多维分析模型的通用方法。该方法将转换规则与实现进行了分离,具有通用性与开放性。

关键词 多维模型,扩展商业报告语言,通用仓库元模型

中图法分类号 TP311

文献标识码 A

Building Multidimensional Analysis Model for Business Reports Based on CWM

PIAO Yong^{1,2} WANG Xiu-kun¹ CHEN Zhi-kui² WANG Zheng²

(School of Electronic and Information Engineering, Dalian University of Technology, Dalian 116023, China)¹

(Software School, Dalian University of Technology, Dalian 116620, China)²

Abstract With standardization and normalization process of financial reports, large numbers of enterprises are introducing extensible business report language(XBRL) to disclose and manage their financial data. Because of complexity and domain specific feature of XBRL, it lacks semantic description for data analysis, leading to the situation that many existing tools capable of multidimensional data analysis cannot understand and deal with XBRL files. Based on the analysis on XBRL structure, the paper applied Common Warehouse Metamodel(CWM), which is platform and domain independent, to build a generic multidimensional analysis model from business reports in XBRL format. Transforming rules are separated from implementation and hence making the model more generic and open.

Keywords Multidimensional model, Extensible business reporting language, Common warehouse metamodel

1 引言

随着企业竞争的国际化与标准化,越来越多的企业使用国际标准 XBRL 进行财务数据的管理与报告^[1]。XBRL 在构成上将实例数据与分类模式分离,可以保证实例数据能够使用不同分类模式进行控制,同时提供了实例数据文件单独安全性控制的可能性。另外, XBRL 基于 XML 技术,提供了一种轻量化级的数据存储和管理方式。

XBRL 的最大问题在于,对于很多企业来说,现有的 OLAP 与数据挖掘工具在数据结构上不能与 XBRL 兼容,这对于应用和信息的集成存在着无法逾越的障碍。一种解决方案是使用“桥”的技术^[2],即在不同的软件系统之间建立特有的数据格式转换工具。这样做的问题是需要建立的桥的数量很多,而且当某数据结构发生变化,需要修改的地方过多,造成维护困难。

为了能够描述更多的财务信息, XBRL 在结构上具有一定的复杂性,这是不利于使用“桥”技术的原因之一。同时, XBRL 所用的主要技术 XLink 不适合进行更多语义信息的表达也是不容忽视的问题。如何将更多的语义信息(如本体信

息)与使用 XBRL 格式描述的财务数据进行融合,也是目前研究的热点之一^[3,4]。

对于数据仓库与挖掘领域, Common Warehouse Metamodel(CWM)^[5]是解决上述问题更为理想的途径。CWM 是一个被 OMG(对象管理组织)采纳为在数据仓库和业务分析环境中进行元数据交换的标准,它完整地描述了数据仓库和业务分析领域的各个方面。CWM 提供构建数据仓库领域模型所需的语法和语义,可以描述一个完整的数据仓库系统的所有组成部分。利用 CWM 元模型(即 CWM 规范),我们可以共享和利用基于 XBRL 的商业报告与各种分析工具。CWM 是一个特定于数据仓库领域的元数据模型集,它独立于具体数据仓库的实现,包含了各类有代表性的数据仓库元数据的有效描述。因此,基于 CWM 的数据仓库和业务分析工具可以理解各类描述特定元数据实例的模型,并可在无需知道彼此专有信息结构和接口的前提下在元数据级别上进行集成。文献[6]针对工程数据仓库系统的异构及分布特点对 CWM 进行了扩充,添加了针对工程数据仓库的元模型,达到一种规范集成的目的。文献[7]提出了一套基于 CWM 的元数据管理方案,适用于本地和分布式环境下的元数据管理,形

到稿日期:2010-10-27 返修日期:2011-02-26 本文受国家自然科学基金(青年)项目(70801008)资助。

朴 勇(1975—),男,博士生,副教授,主要研究方向为智能计算与信息处理, E-mail: piaoy@dut.edu.cn; 王秀坤(1945—),女,教授,博士生导师,主要研究方向为数据库与决策支持系统。

成一个针对元数据的管理系统。文献[8]提出了一种基于 CWM 的结构化数据信息集成架构,主要思想是使用统一的集成元模型进行 Wrapper 设计,使得多异构数据源的情况下只使用单一的 Wrapper 设计就能达到要求,但文章主要针对的是关系型数据源,而且集成的结果主要用于数据的格式显示。文献[9]是基于 CWM 和 EMF 的数据库元数据处理,自动完成领域模型的实现和表结构的生成,但主要也是针对关系型的数据源。

本文以基于 XBRL 的国际财务报告标准(IFRS)为例,给出一种基于 CWM 的商业报告多维分析模型的构建方法。第 2 节简单介绍了 XBRL 技术,并以此为基础在第 3 节给出了 XBRL 数据基于 CWM 的构建模型和系统框架;最后做了总结。

2 可扩展的商业报告语言(XBRL)

XBRL 在商业报告中的优势和作用:(1) XBRL 可以提升商业业务报表信息的生成、交换和比较能力。(2) 计算机软件可以通过识别 XBRL 的标记自动处理报告信息,从而省去了人工输入及对比的高昂人力成本和时间成本。(3) XBRL 可以利用语意信息、元数据和业务规则来自动验证数据的正确性,从而降低了数据错误的风险,减少了审计所需时间。

XBRL 架构将业务报告信息分为两个部分:实例(instance)文件和分类(taxonomy)文件。分类文件又分为两个部分:模式(Schema)文件和链接库(Linkbase)文件,如图 1 所示。

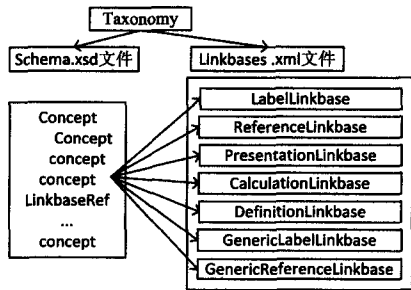


图 1 XBRL 分类机制

实例文档包含被报告的事实数据,分类文件中的模式文件定义表达事实的概念,链接库文件通过定义概念内部的关系以及概念和文档间的关联关系来进一步表达概念的含意。实例文件、模式文件以及链接库文件一起构成了 XBRL 的业务报告。

如图 1 所示,XBRL 分类机制在模式文件中定义了一组概念,然后通过模式文件中的 LinkbaseRef 元素来指向所支持的链接库文件,由这些链接库文件给出概念之间的附加关系,如定义链接库文件中所包含的元素之间的层次结构关系可以包括一般与特殊的关系、原名与别名关系等等。除此之外,通用分类标准使用定义链接库来表达维度信息。

标签(Label)链接库表示元素及其显示名称之间的对应关系,能够将元素与人们更容易理解的名称联系起来。分类机制同时使用某国语言、英文定义元素标签,如:中国证监会使用的就是中文和英文来定义元素标签。

引用(Reference)链接库包含分类中的概念和对给出这个概念含义的商业、经济和会计出版文献中权威陈述的引用之间的关系。

显示(Presentation)链接库描述模式定义文件中所包含的元素间在展现结构上的层次关系和顺序,通常这种展现顺

序与定义链接库中描述的层次结构关系类似。

计算(Calculation)链接库描述模式定义文件中所包含的元素之间的数值计算关系。如总资产=固定资产+流动资产。

通用标签(Genericlabel)链接库为扩展连接角色(ELR)提供定义标签,这些标签是除了用英文表示的其他标签。如元素 IAS_1_2010-04-28-role-210000,其中一个标签为 Bilanz, kurz-/langfristig- Konzernabschluss,使用德语表示。

针对 XBRL 使用者和软件开发商的要求,XBRL International 尽量保证 XBRL 2.1 规范的稳定性,这也是 2.1 版技术规范发布以来没有出现后续版本的原因之一。XBRL 技术规范需要稳定,但是一些功能的需求还需要添加。XBRL International 采取的办法是发布可选用的外加规范模块,既不损害软件开发商的利益,又兼顾 XBRL 发展的需要。XBRL Dimensions 规范就是因此而产生的。

XBRL Dimensions 是 XBRL 2.1 规范的可选功能模块,这个全球标准用来表示在业务数据中常见的多维信息,以便使计算机应用程序可以直接读取这些关系代表的含义。Dimensions 规范提供了表达 XBRL 背景(context)之间关系的能力,它利用 domain-member, hypercube-dimension, dimension-domain 以及 dimension-default 等关联来描述多维信息。遵守 Dimensions 规范的实例文档可以更方便进行数据的深加工,如数据挖掘、数据分析等工作。

3 基于 CWM 的多维财务数据模型

XBRL 元模型的描述是通过 XML 模式文件及 XLink 技术来实现的,而且主要是针对财务报告格式的规范,缺少关于数据对象及其关系的描述,因此 XBRL 模型语义描述能力不强,对于较复杂的多维数据模型,如数据立方体模型或数据仓库中的星型模型等不能很好地建模。也正是这个原因,XBRL 规范主要是用来进行财务数据的展现等显示工作。如果要使用已有的数据仓库、OLAP、数据挖掘等平台做数据的分析工作,必须首先解决基于 XBRL 格式的商业报告模型到多维分析模型的转换,即要使用一种能够识别目前商业智能标准的元模型。

CWM 就是一种自然而然的解决方法。CWM 使用包机制来组织元模型,这种组织结构使其既易于理解、使用(用户建模),又便于裁减和扩展(工具开发者实现)。CWM 目前的版本中包含了 18 个包和底层的 Object Model,每个包尽量涉及一个独立的领域。虽然各个包所描述的领域不尽相同,但它们之间有着紧密的依赖关系。事实上,所有的包按功能和抽象层次组织成 4 层,同层的包的功能角色类似,每一层中的包都为同层或上层的包提供服务,层次越高,描述的内容越抽象。其中的每个包都由一系列 UML 表示的类图来描述,这样极大地方便了系统开发者的建模工作。与多维模型构建关系比较密切的是 CWM 中的 OLAP 包和 Transformation 包,它们分别定义了数据多维分析和转换的基本结构。我们在模型的研究中主要涉及这两个包,但也会使用到底层描述中的定义。

模型的构建首先需要一种将财务报告中的信息抽取出来并按照 OLAP 模型进行组织的一种方法。该方法可以理解 XBRL 的元模型并使用一种遵照 UML 规范的方式进行描述,具有较强的语义表达能力,全面描述了 OLAP 的基本类及其关系从 XBRL 到 CWM 的映射,使得在不同分析平台都

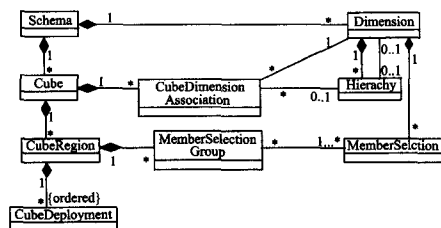
可能进行数据及模型的交换。

这里我们只是给出 XBRL 报告中关于各财务科目项与 CWM 的对应关系,其基本的思想是将 XBRL 报告相关部分抽取出来,并产生对应的数据立方体模型。XBRL 实例文件中的数据会被映射到数据立方体模型中的事实数据;同时,实例文件中与事实数据相关的 Context 等信息会同时被映射为时间、公司分类等维度及其域值。XBRL 中的可发现分类集(DTS)也会被映射为不同的分析模型的维度。其中最值得关注的是定义链接库与计算链接库,因为定义链接库中定义了概念之间丰富的联系,如概念之间的层次关系、一般特殊关系等,实际上其已经包含了显示链接库的所有父子关系,因此对应着多维模型中维度及维度域的层次信息。

在最新标准的 XBRL 规范中,定义链接库开始引入了维度的标准,而且 IFRS 及 GAAP 已经开始应用这种标准。虽然这使得 XBRL 开始具有一定的多维描述,但对于目前的分析应用的集成及财务报告的集成都没有太大的帮助,而且目前其普及程度还比较低。

计算链接库中描述的财务科目项之间存在的某种数据计算关系,如某会计项的计算方法,这本是用作商业报告正确性的检验机制,但从数据的角度来看,这实际定义了多维模型中域成员间的聚合关系,这种关系使得多个数据项联合成为一个数据项。

在 CWM 的 OLAP 包中(见图 2), Schema 元素是包的根元素, 包含 Cube 和 Dimension。Dimension 可以理解为多维结构中的坐标, 由若干不同的值组成, 这些值构成了 Dimension 中具有相同语义的成员域值。Cube 表示在相同维度下分析数据的集合。CubeDimensionAssociation 定义了 Cube 与 Dimension 之间的关联关系。维度可以包含层次 Hierachy, 层次允许 Dimension 中的成员域值可以根据它们之间的父子关系进行遍历, 对应着 OLAP 操作中的上卷或下钻操作。遍历的方式可以是基于导航式的或者计算式的。MemberSelection 选择了 Dimension 中的一部分域值的集合, 这对应着 OLAP 的切片或分块的操作。CubeRegion 是 Cube 的一个子集并且含有 MemberSelection 中的维度, 它对应着一种 Cube 的物理实现, 如 OLAP 中的物化数据。CubeRegion 可以拥有多个 CubeDeployment, 它们对应着多维模型的实现策略。

图 2 CWM 中的 OLAP 结构^[5]

XBRL 链接库的逻辑结构如图 3 所示。这里我们主要考虑定义链接库和计算链接库的结构,该图使用 UML 进行描述。图中 MonetaryItem 是实例文件中事实数据的抽象,它们对应着 XBRL 模式文件中定义的概念。这些概念由不同的 XLink 指针引用,形成了多维模型的立方体和维度的实现。定义链接库中的 DefinitionLinkArc 描述了维度的结构, XBRL 主要是通过属性 arcrole 来具体化各财务项之间的关系,如 Dimension-hypercube 描述了两个概念是维度与立方体的关系,这种关系可以直接映射到 OLAP 中的 CubeDimensionAssociation; Dimension-domain 表示两个概念是维度与域

的关系,除此之外还有 Domain-member, Dimension-default 关系。更一般地,还有 general-specific 等关系,用来描述财务项之间的层次结构(父子)关系。同样,计算链接库中描述财务项之间聚合的关系对应着维度域成员间的逻辑关系。

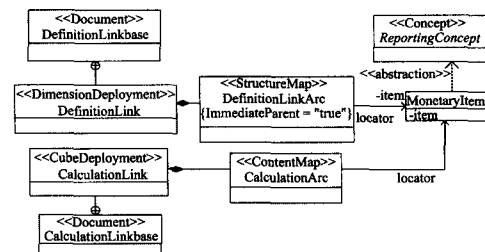


图 3 定义链接库的结构与多维模型映射

为了能够描述 XBRL 到 CWM 结构的映射关系,我们在图 3 中使用了 CWM 中定义的构造型。

4 系统实现的框架

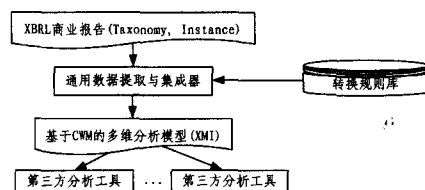


图 4 模型构建过程

```
namespace d = "http://www.xbrl.org/2003/linkbase"
namespace n = "http://www.w3.org/2001/XMLSchema"
namespace xsi = "http://www.w3.org/2001/XMLSchema-instance"
namespace xlink = "http://www.w3.org/1999/xlink"
namespace xbrli = "http://www.xbrl.org/2003/instance"
```

```

let $ds2 := document("commons.schema.xsd")/n:element
let $ds3 := document("schema.xsd")/n:element
for $ds1 in document("definition.xml")//d:definitionLink
let $src := $ds1/d:definitionLinkArc
let $locs := $ds1/d:loc
return
<hierarchies>
{$ds1/@xlink:role}
{ for $src in $src
  let $fromloc := $locs[@xlink:label=$src/@xlink:from]
  let $toLoc := $locs[@xlink:label=$src/@xlink:to]
  let $elems1 := $ds2[@id=substring-after($fromloc/@xlink:href,"#")]
  let $elems2 := $ds3[@id=substring-after($fromloc/@xlink:href,"#")]
  return
  <arc>
  { IF (count($elems1) > "0") THEN
    <data1>
    { for $e1 in $elems1
      let $eto11 := $ds2[@id=substring-after($toLoc/@xlink:href,"#")]
      let $eto12 := $ds3[@id=substring-after($toLoc/@xlink:href,"#")]
      return
      <data2>
      { IF (count($eto11) = "0" and count($eto12) > "0") THEN
        <h><parent>
        {$e1/@id}</parent><child>{$eto12/@id}</child>{$src/@order}
        </h>
        ELSE IF (count($eto11) > "0" and count($eto12) = "0") THEN
        <h><parent>
        {$e1/@id}</parent><child>{$eto11/@id}</child>{$src/@order}
        </h>
      }
    }
    </data2>
  }
  </data1>
  .....
}
</arc>
}
</hierarchies>

```

图 5 一个转换规则的示例

上节中的映射关系我们通过 XQuery 语言进行了实现,并保存于转换规则库中。XQuery 的解析是通过解析器解释运行的,其好处是使规则与系统分离,便于以后的维护。提取出来的商业报告数据按照 CWM 定义的格式进行集成,最后形成基于 CWM 的多维分析模型,并以 XMI(XML Metamod-

el Interchange)的格式导出。此 XMI 文件可以被理解的第三方分析工具导入并进行进一步的多维展示和操作。图 4 给出模型构建过程。

图 5 给出了一个转化规则的示例片段。这段代码主要是从定义链接库中按照 DefinitionLinkArc 的描述,整理概念之间的父子层次关系,形成维度及其层次结构,这是构成图 3 中构造型<<StructureMap>>的一部分。

结束语 本文给出了一种基于 CWM 的财务数据多维分析模型的构建方法,通过此模型可以将缺少必要语义描述的 XBRL 财务数据报告进行转化和集成,生成由商业智能标准描述的多维模型。这多维模型可以被工业界成熟的其它数据分析工具,如数据仓库、OLAP 分析等所识别并进一步进行数据的多维展现和分析,使得基于 XBRL 商业报告的多维分析更为方便。本系统采用 XQuery 描述模型的构建规则并统一管理,因此具有较好的通用性及开放性。

参考文献

- [1] XBRL 国际组织. Extensible Business Reporting Language (XBRL) 2.1[S]. 2008
- [2] Poole J, Mellor D, Chang D, et al. Common Warehouse Metamodel[M]. Wiley, John & Sons, 2001

(上接第 163 页)

PUBLISHING SECURE VIEW(S, p_v, V', θ)

输入:隐私查询 $S(IA:ia, SA:sa)$,已发布视图集 V 的隐私泄露概率 p_v ,待发布视图集 V' ,隐私泄露阈值 θ ;

输出:相对安全的视图集 V' 。

Step1 初始化 $p=0; p_{max}=p_v$;

Step2 /* 视图集 $V \cup V'$ 的隐私泄露概率计算 */

$V_k = \{ V_i \mid V_i \in V', ia \in V_i \};$

$V_l = \{ V_j \mid V_j \in V, sa \in V_j \};$

if $V_k \neq \Phi$ and $V_l \neq \Phi$ then $p = \text{PRIVACY DISCLOURE}(S, V_k, V_l);$

if $p > p_{max}$ then $p_{max} = p$;

for each V_i where $V_i \in V$ and $ia \in V_i$, do

{ $p = \text{PRIVACY DISCLOURE}(S, V_i, V_l);$

if $p > p_{max}$ then $p_{max} = p$;

for each V_j where $V_j \in V$ and $sa \in V_j$, do

{ $p = \text{PRIVACY DISCLOURE}(S, V_k, V_j);$

if $p > p_{max}$ then $p_{max} = p$;

Step3 Leak = $p_{max} - p_v$; /* Leak 值计算 */

if Leak $\leq \theta$ then

{发布视图集 V' ;

$p_v = p_{max}$;

else

{删除 V' 中的 ia 和 sa ;

发布视图集 V' };

Step4 return(p_v)。

结束语 本文从相对安全角度出发,利用条件概率思想,综合考虑了视图间的联系和制约关系,研究了动态视图的隐私泄露问题。本文的视图由一般的 PSJ 查询表达式生成,具有较好的实用性。提出的可能世界构造方法使隐私泄露概率计算更容易实现,也更具有实际意义。后期工作将探索隐私查询其他表示时的隐私泄露概率计算公式,使该公式更具一般意义。

参考文献

- [1] Sweeney L. K-Anonymity: A model for protecting privacy [J].

- [3] Lara R, Cantador I, Castells P. XBRL taxonomies and OWL ontologies for investment funds[C]// Tucson A Z. ed. 25th International Conference on Conceptual Modeling(ER 2006). Springer Verlag, 2006
- [4] Spies M. An ontology modelling perspective on business reporting[J]. Information Systems, 2010, 35(4): 404-416
- [5] OMG. Common Warehouse Metamodel Specification [S]. 2003
- [6] 赵晓非, 黄志球, 沈国华, 等. 基于元模型的工程数据仓库系统元数据集成[J]. 南京航空航天大学学报, 2006, 38(3): 341-346
- [7] 田耕, 宁洪, 李姗姗, 等. CWM 研究及相关元数据管理系统的设计[J]. 计算机工程, 2006, 32(11): 100-102
- [8] 夏秀峰, 孙娜, 石祥滨, 等. 基于 CWM 的结构化异构数据抽取方法研究与实现[J]. 计算机应用与软件, 2009, 26(12): 108-110
- [9] 陈兴建, 郝文宁, 靳大尉, 等. 基于 CWM 和 EMF 的数据库元数据处理[J]. 计算机工程, 2010, 36(13): 40-41
- [10] Liu Z H, Baby T, Krishnamurthy S, et al. XBRL repository-An industrial approach of management of XBRL documents[C]// 26th IEEE International Conference on Data Engineering(ICDE 2010). IEEE Computer Society, 2010: 1037-1047
- [11] Ma J, Pan X, Yao J. The research on the knowledge representation model of XBRL[C]// IEEE International Conference on Service Operations and Logistics, and Informatics (IEEE/SOLI 2008). IEEE Computer Society, 2008

In'tl Journal on Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(5): 557-570

- [2] Dalenius T, Reiss S. Data swapping: A technique for disclosure control [J]. Journal of Statistical Planning and Inference, 1982, 6(1): 73-85
- [3] Halevy A. Answering Queries Using Views: A survey [J]. VLDB Journal, 2001, 10(4): 270-294
- [4] Bancilhon F, Spyrators N. Protection of Information in Relational Data Bases[C]// VLDB. Tokyo, Japan, 1977: 494-500
- [5] Samarati P, Sweeney L. Protecting Privacy When Disclosing Information: k-anonymity and Its Enforcement through Generalization and Suppression[R]. SRL-CSL-98-04. SRI Computer Science Laboratory, 1998
- [6] Miklau G, Suciu D. A Formal Analysis of Information Disclosure in Data Exchange[C]// Proceedings of the 20th ACM SIGMOD International Conference on Management of Data. Orlando, USA, 2004: 507-534
- [7] Dalvi N, Miklau G, Suciu D. Asymptotic Conditional Probabilities for Conjunctive Queries [C]// Proceedings of the Sixth International Conference on Database Theory. Edinburgh, UK, 2002: 289-305
- [8] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression [J]. Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 571-578
- [9] Sweeney L. K-anonymity: A Model for Protecting Privacy[J]. Int'l Journal on Uncertainty, Fuzziness, and Knowledgebased Systems, 2002, 10(5): 557-570
- [10] Yao Chao, Wang Xiao-yang Sean, et al. Checking for K-Anonymity Violation by Views[C]// VLDB. 2005: 910-921
- [11] Li Ning-hui, Li Tian-cheng, Venkatasubramanian S. t-Closeness: Privacy Beyond k-Anonymity and l-diversity[C]// Proceedings of the 23th International Conference on Data Engineering. Istanbul, Turkey, 2007: 106-115