多维数据的在线自组织方法①

An Online Self-Organization Method on Multidimensional Data

姜　婷　肖　刚　高　飞　陆佳炜　(浙江工业大学 计算机应用 浙江 杭州 310014)

摘　要： 随着 Internet 的高速发展,用户对在线服务质量的要求越来越高，本文提出了一种在现有硬件资源的基础之上提高数据在线组织管理灵活性的方法。指出了数据的多维度、结构化的特点，采用用户请求的异步通信方式，结合动态树视图实现多维数据的在线自组织，实验结果证明了这种方法在数据自组织管理、web 响应延迟等诸方面均能取得令人满意的效果。

关键字： 多维数据 自组织分类 动态树视图 异步通信

1　引言

随着 Internet 的发展及关系数据库系统技术成熟，大量企业数据存放在关系数据库中，通过网络进行管理。针对用户对于在线服务质量的要求越来越高，如何在现有硬件资源的基础之上，分析出数据中有效的、隐含的、有潜在使用价值的相关信息将有助于企业的可持续发展。

数据分类(classification)是知识发现领域中最重要的问题之一，旨在发现属于同一类数据对象的共同特性，构造分类器[1]。目前，已有若干种方法和技术用于构造分类模型，如文献[2,3]采用决策树，是一种自顶向下的贪心算法，在每个结点选择分类效果最好的属性，但当属性值较多时，效果可能就会比较差；文献[4-6]采用 P2P 网络,基于分布式 Hash 表的方法当簇间通讯量增加时，网络的性能随之降低。影响数据在线组织的有效性除了分类方式外，还需要考虑 Web 服务响应时间[7]。目前对于 Web 性能进行优化使用最广泛的技术是动态缓存(dynamic caching) [8]，但如果数据的组织形式发生巨大改变，缓存中的大量页面将被标注为失效；相应地，缓存的命中率将大大地降低，随之而来的是脚本重新生成大量的数据传输占用网络宽带。

本文以 web 应用为背景研究数据的在线自组织，将存储在数据库中的数据信息看成是用多维空间描述的多维数据，将它结构化的多维度特点作为分类层次的依据，用动态树视图方式实现多维数据的自组织分类，用户能按需对多维数据实现分类，采用用户请求的异步通信方式，减轻网络通信流量负荷、提高系统的响应时间。

2　多维数据的自组织分类

数据的组织分类是利用数据间隐含的语义关系数据源进行高效组织。帮助人们从不同的角度对所关心的事实进行分析，从而为各级管理人员提供辅助决策信息。大多数应用系统的数据都是以二维表的形式存储在数据库中，将表中的属性作为维度，用维空间描述数据源。不同的属性组合构成不同的维空间，所以维空间对数据源的描述不是唯一的。多维数据的自组织分类是通过构造不同的维空间来实现数据组织结构的动态变化。

在实际分析过程中，分析人员并不是将所有维度纳入到分析活动中，而是选择具有密切关系的维度集纳入到分析活动中。在数据集 D 的属性集{A1,…,An}中选取属性 A1、A3、A5 作为维度集构成三维空间，以维度 A1 为前提，数据集 D 被划分成若干子集并形成新的数据集 D1 映射在 A3A5 空间；同理以维度 A3 为前提，数据集 D1 映射为数据集 D2；以维度 A5 为前提，数据集 D2 映射为数据集 D3 分布在 A5 轴上，

① 基金项目:浙江省自然科学基金项目(Y106603)
收稿时间:2008-09-02

数据集 D 随着逐级映射实现了分类。

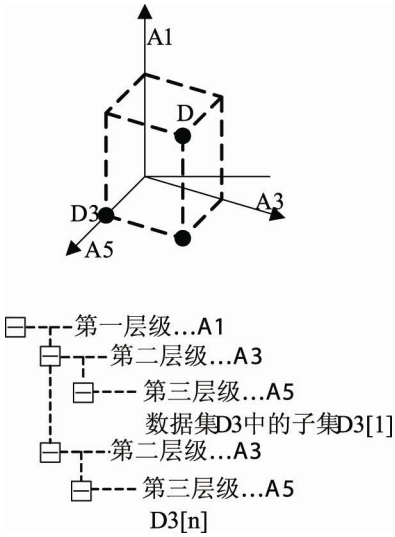

图 1　多维空间映射与树视图的对应关系

把树视图空间解释为维度空间，树视图的层次解释为维度，树视图的某一层级对应于某一维度上的映射，数据在维空间上的分类可以由树视图模型映射出来。例如，图 1(b)用以 A1 作为第一层级，以 A2 作为第二层级，以 A3 作为第三层级的树视图描述数据集 D 的分类，数据集 D3 中的子集就对应于树视图第三层级上的数据。数据信息作为多维空间的一个实体，数据信息的属性就是多维空间的数据维，树视图空间就是高维空间的一个子空间，数据按属性动态分类可以看作是数据从高维空间向低维空间的映射。

3　多维数据在线自组织模型

本文提出了一个在线自组织树模型描述多维数据在线自组织。在线自组织树模型由两个问题组成：数据组织、信息展示。数据组织主要是指将数据从高维空间向低维空间的映射。由维度的组织信息确定树的层次结构，树视图的每一次动态生对应着不同的数据组织结构，实现了动态定制树视图层次结构也就实现了在不同维度下多维数据自组织映射。信息展示是指将数据组织形式可视化，反映在树视图上就是将组织好的信息以树分类层次的形式显示出来。在多维数据自组织的树视图中，映射维度的个数对应树的层次数，每个维度对应树的层次；维度间的偏序性对应树的组织结构。

在线自组织树模型如图 2 所示，服务器端实现数据组织，客户端实现信息展示。自组织过程是经由路径分析器分析维度的组织特征，作为树层次结构自组织运算的输入条件，通过树层次预处理器动态生成树形框架，数据映射引擎就可以根据树的层次框架，实现对数据逐级映射。该模型不仅提高服务器端系统响应，还考虑到从数据传输和页面响应速度，在服务器与客户端之间采用基于用户请求的异步通讯方式进行数据的交互。

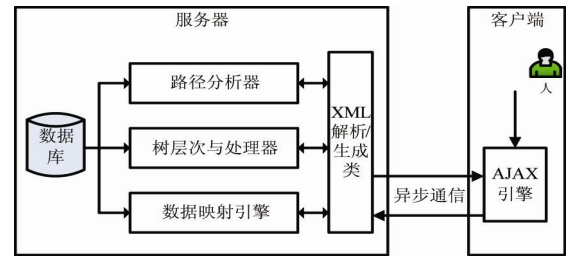

图 2　多维数据自组织的书试图模型结构

4　运行时在线自组织的动态交互与异步通信

4.1 树视图运行时的动态交互特性

自组织的思想是指在运行时通过动态生成分类层次结构来实现数据的自组织。该思想在树视图中体现为树视图运行时动态交互特性，改变树视图层次结构，从而改变数据组织组织结构。

由 1 节可知，树视图空间不是唯一的，依据实际需求，用户通过规则对维空间描述进行不同的映射，得到不同的树视图空间结构，从而实现数据组织结构的变化。维空间特征定义语法如下：

< Space >::=< SpaceName > |< dimension >|, |< dimension >|

|

< dimension >::=< dimensionname > |

location: < P_ location>

order: < dimensionorder >

|

<dimensionorder>::=<dimensionname>→<dimensionname>

维度的组织特征由维度名、物理位置、偏序关系组成。路径分析器根据规则将维空间特征转换成“属性——值”的 XML 文件形式，XML 文件结构为：

<database>

<dimension 1><row name=”” /><row location= ”” /><row order= ”” /> …

</dimension 1>

…

<dimension n><row name=""/><row location=""/><row order=""/>…</dimension n>

</database>

XML 文件中 <dimension1> 到 </dimension1>是整个 DOM 树的一个分支，存储了对某一维度组织特征的描述信息。其中，dimension 1..n 表示维度的可变属性，采用这种形式的好处在于在 XML 解析类中不会出现维度的名字，可以用抽象的函数遍历各个节点。

通过对 XML 描述的维度组织特征进行学习，使静态的描述维度组织特征的数据成为一种动态的数据参与到多维数据分类的自组织运算。将维度依其偏序性映射到树视图，偏序性越靠前的维度，映射的层次越高。然后，将数据以学习生成的映射路径进行聚类运算。运行时期，当客户端需要变更数据的组织结构时，因为树视图空间的映射依据维空间特征所描述的寓意，所以利用元操作修改维度信息，就能动态地调整树视图的层次结构。

为了提高系统响应速度，自组织树模型在对数据进行逐级映射之前需要进行预处理，先构造树的层次框架，而不是直接对数据进行分类。树的层次框架只有层次结构，没有数据。利用 XML 文档存储层次结构特征描述，XML 的描述如下：<leveli id='' name='' value='' child='' open='' >树的第 i 层对应标签 leveli，其具有四个基本的属性：id，本级名称，值，子级个数，是否打开。XML 灵活的读写能迅速的定位到相应的层级，为实现数据映射提供方便，提高运算的灵活性和扩展性。

4.2 基于用户请求的异步通信

当数据量变大时，即使利用预处理定义好的树层次对数据进行分类，后台的响应时间也会变得迟缓，客户端脚本响应超时。引入基于用户请求的异步通信思想，即能解决对大量数据分类的耗时，又能减轻网络的传输负担。

在传统的 Web 应用的工作模式下，基于 WEB 的应用系统总是让用户进入“提交→等待→重新显示页面”的一个传统循环，用户的操作总是要等待服务器的响应，本文利用 AJAX 提供与服务器异步通信的能力，能让用户从传统的请求/响应模式中解脱出来。AJAX 能减轻服务器的负担；无刷新地更新页面，减少用户的等待时间。借助于 AJAX，可以让用户在操作时使用包含在 HTML 中的 JavaScript 脚本向服务器发出异步数据请求，服务器返回一个只包含 XML 格式数据的响应，并根据所得数据使用 DOM(文档对象模型)生成或更新客户端用户界面。

基于用户请求的异步通信的自组织树视图，初始化时不需要对所有的数据进行分类，后台只需预处理生成分类层次，只有当用户发出请求时，根据请求确定相应的路径查找相应的数据返回给用户。图 3 为基于用户请求的异步通信工作框架图：

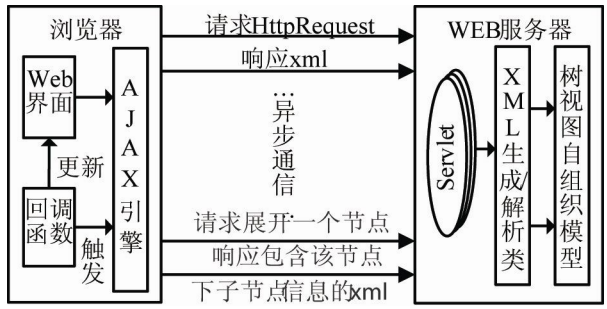

图 3　基于用户请求的异步通信工作框图

客户端与服务器之间的异步通信过程如下：

(1) 前台浏览器捕捉用户的操作事件。

```
If need new data
    creatXMLHttpRequest(value);
```

用户对数据验证和数据处理的请求提交 AJAX 引擎来做，只有确定需要服务器读取新数据时再由 AJAX 引擎代为向服务器提交请求。

(2) 请求是以 XML 文档的形式传输，后台服务器端解析 XML 请求信息，数据处理完成经由 XML 解析/生成类，将处理结果存放在 XML 文档中，通过 response 传输到客户端。

```
PrintWriter out = HttpServletResponse response. getWriter();
out.write(xmlString);
```

(3) 客户端获取服务器端的 XML 文档，动态地编辑页面。

```
if req.readyState is succeed{
        var xmlDoc = new ActiveXObject ("MSxml2.DOMDocument")
        xmlDoc.loadXML(req.responseTe xt);
        add(xmlDoc); }}
```

其中，req 为 XMLHttpRequest 对象，当它的 readyState 属性改变时触发 complete()事件处理句柄，当 req.readyState 表示成功返回数据，通过：

element=document.createElement(tagname);

element.setAttribute(attribute,value);

Pelement.appendChild(element);

创建元素，给新元素添加各种属性，并将元素添加到页面中已存在的 Pelement 元素下，作为 Pelement 元素下的子节点，完成网页的动态添加。

AJAX 技术使用户操作与服务器响应异步化。将递归放在服务器端实现，利用 XML 中标签的层次性描述数据的树形组织结果，使得 JavaScript 脚本被下载到客户端后按顺序的方式描绘树形。采用用户请求的方式动态加载数据，可以减轻脚本运行的负担，每次响应 xmlHttp 从服务器端传回的只是用户需要的数据。

5　实例

这种多维数据在线自组织方法已在浙江人事厅专家信息管理系统中得到很好的应用。用户通过树视图定制模块根据需求自定义树的层次，实现对专家基本信息进行按需的自组织分类。实例中，当前台用户欲改变数据的组织结构，如评委会作第一层，专业组作第二层，专业组职务作为第三层属性，后台就会响应这种改变，动态生成树的层次框架，返回给用户的是有关第一层节点相关信息的 xml 文档。

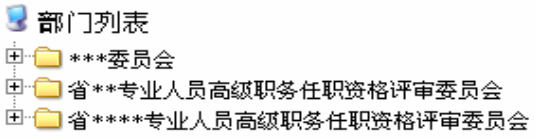

图 4　初始化节点

用户首先需要浏览的是树视图的第一层次如图 4。如果用户请求展开“***委员会”节点，将该节点的路径作为参数传递到服务器端。在树视图层次框架 xml 文档中定位到相应的节点并查找相关数据，将数据组织结果以 XML 文档中形式返回，异步请求返回的 XML 文档如下：

<return id="1" list="5" name="level2">

<level2 id="1|1" name="隶属专业组名称" value="01" list="1" open="0" value1="哲学" />

…

<level2 id="1|1" name="隶属专业组名称" value="01" list="1" open="0" value1="英语组" />

</return>

解析服务器端返回的节点参数，定位到触发事件的节点，动态地编辑页面。如图 5：

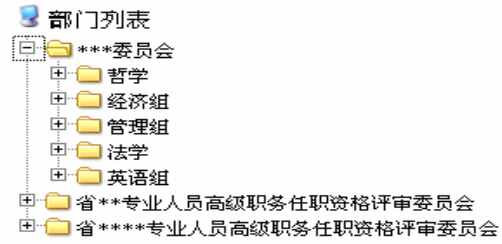

图 5　异步请求下层节点信息

6　结束语

这种在线自组织方法能在运行时动态改变数据的组织形式，用户可以根据不同的需求灵活地管理大量的信息数据。用户请求的异步通信，使系统在初始化时，不需要将所有的数据都进行传输而是按需传输，该方法的优点就是将不变的总传输量离散化，用户请求什么数据就加载什么数据，请求数据操作即不打断用户的操作,又具有迅速的响应能力，带来极好的用户体验。

参考文献

1 彭京,唐常杰,元昌安,李川,胡建军.一种基于概念相似度的数据分类方法.软件学报,2007,18(2):311－322.

2 郑向群,赵政.基于 S-CART 决策树的多关系空间数据挖掘方法.计算机应用,2008,28 (3):749－752.

3 魏晓云.决策树分类方法研究.计算机系统应用,2007,9(2):42－45.

4 Crespo A, Garcia, Molina H. Semantic overlay networks for P2P systems. Technical Report, Stanford University, 2002.

5 Loeser A.Taxonomy-based overlay networks for P2P systems. Proceedings of IDEA S, 2004.

6 乔百友,王国仁. Super_Peer 网络中基于语义的分簇算法研究.小型微型计算机系统, 2008,29(2):213－218.

7 王亚沙,赵俊峰,谢冰.基于用户视角的组合 Web 服务响应时间优化.计算机学报, 2006,29(7):1179－1188.

8 凌波,王晓宇,周傲英,Ng Wee-Siong.一种基于 Peer_to_ Peer 技术的 Web 缓存共享系统研究.计算机学报, 2005,28(2):170－178.

9 Prigogine L,Nicolis G. Self-Organization in Nonequilib-rium Systems. John Wiley & Sons,1977－1989.

10 董攀,朱培,卢锡城. 一种网络的自组织演化的数学模型.软件学报, 2007,18 (12):3071－3079.