<<XML挖掘>>

图书基本信息

书名：<<XML挖掘>>

13位ISBN编号：9787308102544

10位ISBN编号：7308102548

出版时间：2012-7

出版时间：浙江大学出版社

作者：潘有能

页数：152

字数：196000

<<XML挖掘>>

内容概要

潘有能编著的《XML挖掘》内容分为8章，第1章先对XML和数据挖掘技术作简要介绍。
在对XML文档进行挖掘之前，需要先进行文档解析及文档标记消歧，即为第2章的内容。
第3章和第4章分别介绍XML挖掘的两项主要功能：聚类与分类。
和HTML中的超链接一样，XML文档之间也具有链接性，第5章介绍利用链接挖掘XML文档间结构的方法。
针对XML文档的查询、检索以及信息提取有利于用户准确、快速、有效地利用XML文档，本书的第6章即讨论XML查询与信息提取技术；第7章和第8章则介绍基于XML数据挖掘建模、知识表示以及Web日志挖掘。

作者简介

<<XML挖掘>>

书籍目录

第1章 XML与数据挖掘概述
1.1 XML
1.2 数据挖掘概述
第2章 XML数据预处理
2.1 XML文档解析
2.2 XML文档标记语义消歧
第3章 XML聚类
3.1 XML聚类概述
3.2 XML文档相似度计算
3.3 XML文档聚类
第4章 XML分类
4.1 相关定义
4.2 权重计算
4.3 相似性计算
4.4 XML文档分类
第5章 XML文档间结构挖掘
5.1 XML链接
5.2 Web结构挖掘算法
5.3 基于XML链接的文档间结构挖掘
第6章 XML查询与信息提取
6.1 XML查询语言
6.2 特征提取
6.3 主题提取
6.4 自动摘要
第7章 基于XML的数据挖掘建模和知识表示
7.1 基于XML的数据挖掘建模
7.2 基于XML的知识表示
第8章 基于XML的Web使用挖掘
8.1 基于XML的Web使用挖掘体系结构
8.2 XGMML
8.3 LOGML文档的结构
8.4 LOGML文档的生成
8.5 基于LOGML的数据挖掘
附录一：基于语义的XML文档相似度计算源程序
附录二：XML文档聚类算法源程序
参考文献

章节摘录

（1）有时候聚类结果是次优解，因为在聚类的初始阶段需要指定簇的初始聚类中心或者均值，当选择的初始中心或均值比较接近实际质心或中心点时会大大地减少迭代的次数，而且可以得到理想的聚类结果。

但如果初始聚类点选择不好，这两种算法又都是一种爬山算法，就会很容易得到次优解。

（2）k值的确定，聚类算法在运行前需要事先指定划分聚类的簇的数目，因为聚类算法是无监督的聚类，并不能了解数据本身特征和整体实际分布情况，因此无法给出一个比较理想的聚类数目。

（3）这两种算法只适用于球状这种特定形状的数据，不适合非球状簇。

（4）k—means算法对噪声和离群点的数据是敏感的，因为少量的这类数据能够对均值产生极大的影响。

k一中心点算法虽然可以减少噪声数据和离群点的干扰，但算法复杂度比较高，因为更新簇的中心点代价比更新簇的均值的代价大的多。

在利用基于划分的聚类算法对XML文档进行聚类时，一般采用k—medoids聚类算法，因为XML文档是一个个离散的对象，当采用k—means算法时，均值并能反映整个簇的实际情况。

另外由于k—medoids算法具有划分算法简单、执行时间快的优点，在XML文档聚类中获得了广泛的应用。

3.1.2层次聚类算法 层次聚类算法是将数据对象组成一棵聚类树的过程，根据生成聚类树的过程是合并还是分裂，可以将层次聚类算法分为两种，一种是凝聚层次聚类（AGNES），另一种是分裂层次聚类（DIANA），如图3.3和图3.4所示。

凝聚层次聚类算法大体过程：首先将每个对象都看作为一个簇，然后度量簇间的距离，根据距离的远近逐渐合并簇，直到所有的对象都在同一个簇中或者满足终止条件。

图3.4描述了一个凝聚层次聚类的过程，它展示出对象是如何一步步合并成一个簇的。

在L＝0层时，a、b、c、d、e对象分别为一个簇；在L＝1时，簇a、簇b的相似度为0.8，大于其他簇间的相似度，所以将a、b合并为一个簇｛a、b｝；在L＝2时，簇d、e间的相似度为0.6，大于其他簇间的相似度，所以将d、e合并为一个簇｛d、e｝，以此类推直到合并为一个类簇为止。

分裂层次聚类算法恰好与凝聚层次聚类算法相反，它先将所有的对象都看成同属于一个簇，然后将原来的簇不断划分成越来越小的簇，直到每个对象自成一簇，或者达到了某个终止条件。

比如，达到了希望的聚类的簇的数目，或者达到了簇间相似度或距离的某个阈值。

版权说明