第 42 卷 第 6A 期
2015 年 6 月

计 算 机 科 学
Computer Science

Vol. 42 No. 6A
June 2015

基于关联分析的网络数据可视化技术研究综述

孙秋年 饶 元

(西安交通大学软件学院 西安 710054)

摘 要 当今万维网、社会关系网等网络的规模迅速发展,海量高维的网络论坛数据给论坛管理员和其他分析人员提出了巨大的挑战,人们很难对隐藏着丰富信息资源的网络论坛数据进行管理和分析。关联规则可以挖掘数据中隐藏的关联关系并预测其发展趋势,可视化技术则能将数据清晰直观地展示,辅助用户决策。于是,针对数据量大、结构复杂的网络论坛数据,将关联分析与数据可视化结合,阐述关联规则和网络数据可视化相关定义及总体目标,并对相关技术进行综述,提出了包含关联规则挖掘、主题挖掘、可视化等技术的基于关联分析的网络数据可视化技术实现框架,以帮助人们在有限的时间内快速理解和分析海量论坛数据集。最后,对数据可视化目前存在的问题与挑战进行探讨。

关键词 可视化,关联规则,网络数据,论坛数据,数据挖掘

中图法分类号 TP391 文献标识码 A

Survey of Network Data Visualization Technology Based on Association Analysis

SUN Qiu-nian RAO Yuan

(School of Software, Xi'an Jiaotong University, Xi'an 710054, China)

Abstract With the rapid development of world wide web and social relationship networks, forum administrator and other analysts are confronted with great challenge from massive high-dimensional forum data. It is difficult for people to manage and analyze network data with rich information resources behind it. While, association rules can provide an effective way to find potential relationship and predict its trend, and visualization techniques can help to display data and assist our decisions clearly. By combining the correlation analysis and data visualization, we focused on forums with files of data and complex structure, illustrated related definition of association rules and data visualization and its overall target, and summarized related techniques. We raised an implement framework of network data visualization technology based on association analysis including the association rules mining, topic mining and visualization and other techniques, which can help people understand quickly in the limited time and analyze massive data set from forums. Finally, we discussed the prime problems and challenges existing in data visualization.

Keywords Visualization, Association rule, Network data, Forum data, Data mining

1 引言

随着万维网、社会关系网等网络的迅猛发展,可获得的网络数据规模逐渐增大,以致人们无法通过传统的技术和方法来管理这些数据。论坛以其门槛较低、聚众能力强的特点成为了网络数据主要的组成部分。这些海量的网络论坛数据给论坛管理员和其他分析人员提出了巨大的挑战。

网络论坛数据结构复杂,数据量又大,人们理解起来非常困难。“一图胜千言”这句谚语告诉我们:一张图像传达的信息等同于相当多文字的堆积描述[1]。海量数据无法直接分析,通过可视化,可以更容易、更快速地从中获得想要的知识。可视化技术起源于 20 世纪 80 年代出现的科学计算可视化(Visualization in Scientific Computing),它是指利用计算机图形学、计算机图像处理、计算机信号处理等方法对数据、信息、知识的内在结构进行表达[2]。可视化借助于人眼快速的视觉感知和人脑的智能认知能力,可以起到清晰有效地传达、沟通并辅助数据分析的作用。

进入 21 世纪,单一的可视化已不能满足人们日益增长的对于挖掘数据中存在的关联关系的需求,可视化逐渐发展为一个涉及数据挖掘、人机交互、计算机图形学等的交叉学科。数据挖掘技术可以帮助人们从海量数据中获取有效的信息。于是,将数据挖掘技术与可视化技术相结合起来,利用人类的认知能力来对大型多维数据集进行数据挖掘是目前人们从海量数据中提取信息极其有效的方法。

Card 认为:信息可视化是从原始数据到可视化形式再到人的感知认知系统的可调节的一系列转换过程。目前已经有很多的可视化系统,例如文献[3]设计实现了一种由多条博文的历史传播数据构成的传播网络可视化系统。由 Jiawei Han 等人提出的图的 OLAP 框架为大型网络可视化提供了新的思路:首先展示大型网络的高层次结构,然后对感兴趣的部分进行向下钻取,以获取隐含的更丰富的知识。在此基础上,文献[4]设计实现了一个利用快速社区挖掘算法对网络的结构

孙秋年(1988—),男,硕士生,主要研究方向为关联分析、数据可视化技术,E-mail:330244581sun@163.com;饶 元(1973—),男,副教授,博士生导师,主要研究方向为社会网络条件下的服务计算、数据挖掘。

信息进行实时分析的大型网络可视化系统。

但是，目前依然缺乏具备数据挖掘功能的网络论坛数据可视化系统。于是，针对网络论坛数据，本文试图将关联分析与数据可视化结合，阐述关联规则和网络数据可视化相关定义及总体目标，并对相关技术进行综述，提出基于关联分析的网络数据可视化技术实现框架。最后，对数据可视化目前存在的问题与挑战进行探讨。

2 网络数据可视化的相关定义、目标及实现框架

2.1 关联规则及网络数据可视化的相关定义

2.1.1 网络论坛数据

网络论坛也被称为 BBS(Bulletin Board System)论坛，是大众发表言论、进行思想交流的平台。与其他领域可视化需求的数据集相比，它具有以下几个特点。

(1)数据量大。论坛数量不断地增加，论坛分为多个板块，板块下有成千上万的帖子，比较热门的帖子又有很多的回复。一般的论坛中有上百万个用户，所以整体数据量很大。

(2)数据层次化。在一般的论坛中，信息都以层次的方式组织，层次模型如图 1 所示。

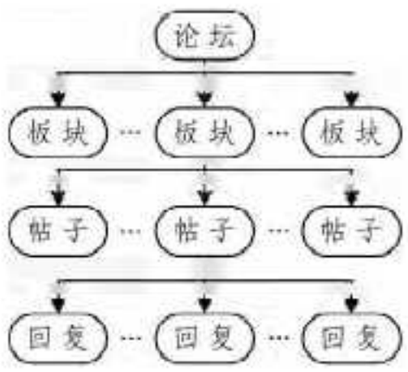

图 1 网络论坛数据的层次模型

(3)交互性高。在论坛上，人们通过发帖或跟帖发表意见，发表的内容任何人任何时间都可以看到。一个帖子往往会引来成百上千的回帖。传播者和受众之间有着灵活的沟通交流机制，即使兴趣不同的网络群体在不同的网络论坛空间中仍然可以相互分享信息、展开讨论，获得彼此的认同。

2.1.2 数据可视化定义

数据可视化可以描述为依据数据的属性特征，借助图形化手段，清晰有效地传达与沟通信息，使通过数据表达的内容更容易被理解。不同数据集的特性不同，选择合适的显示方式和技术，能达到更好地展示出数据本身的结构特征的目的。

2.1.3 关联规则的相关定义

关联规则形式简单，有助于人们发现数据之间的联系和许多其他有趣的模式，因此利用关联规则挖掘网络论坛数据中的关系。有必要先给出关联规则的定义。

定义 1 设 $I=\{I_1,I_2,\cdots,I_m\}$ 是帖子的属性集，称为项(Item)。给定一个存储大量网络论坛数据的数据库 D，其中每个帖子 T 是项的对应数据集合，满足 $T\subseteq I$。每个帖子都有一个标示符，称为 TID。X 是 I 的子集，如果 $X\subseteq T$，则称 T 包含 X；如果 X 的元素个数为 K，则可以称 X 为 K-项集(K-Itemset)。

定义 2 如果项集 $X\subseteq I,Y\subseteq I$，并且 $X\cap Y=\varnothing$，则形如 $X\Rightarrow Y$ 的蕴含式称为关联规则，其中，X 是规则的前项集，Y 是规则的后项集，它表示包含 X 项集的帖子 T 也很有可能会包含 Y 项集。如果包含 X 的帖子有 $c\%$ 也包含 Y，那么规则 $X\Rightarrow Y$ 的置信度为 $c\%$；如果 D 中有 $s\%$ 的帖子包含 $X\cup Y$，那么规则 $X\Rightarrow Y$ 的支持度为 $s\%$，其计算表达式分别为

$$Support(X\Rightarrow Y)=P(X\cup Y) \tag{1}$$

$$Confidence(X\Rightarrow Y)=P(X|Y) \tag{2}$$

2.2 网络论坛数据可视化的目标

相对于传统的可视化模型，基于关联分析的网络论坛数据可视化需要实现以下目标。

(1)网络论坛数据可视化

由于论坛数据量很大，一次性显示全部信息可能导致界面混乱与重叠，不能达到期望的可视化效果，而且需要计算的数据量会比较大，可能导致系统执行时间较长。

(2)挖掘网络论坛数据中隐藏的有价值的信息和规律

由于网络论坛数据交互性高，产生的网络关系复杂，很难发现数据中的规律。在充分利用信息资源，发掘数据中的关联关系和规律方面，关联规则挖掘技术以其可以从数据集中发现属性间隐藏的、有趣的关联关系的优势脱颖而出。

(3)挖掘过程与挖掘结果可视化

数据挖掘和可视化技术的有机结合，可以弥补传统数据挖掘过程的缺陷，加强数据挖掘的处理过程。可视化的方法使数据挖掘技术的应用更具形象性和直观性，挖掘过程加入更多的人的参与和指导，可以有效地提高数据挖掘结果的可信度、可理解性和可用性。可视化贯穿了挖掘过程和挖掘结果，有利于用户对挖掘算法中的参数及时做出调整。

(4)合适的交互操作

设计一种关联规则与其对应数据的交互，用户对某一规则或者某几个规则感兴趣的时候，可以通过下钻找到其对应的数据集，了解数据的详细信息，也可以通过上卷返回到关联规则界面。

(5)快速掌握和了解关联规则下数据的主题思想

主题挖掘可以快速地发现数据集中主要的观点和话题，可以让分析人员快速地对关联规则进行合理的筛选和过滤，选择出人们需要的有意义的规则。

2.3 关联分析的网络数据可视化技术实现框架

关联分析的网络数据可视化技术实现框架如图 2 所示。

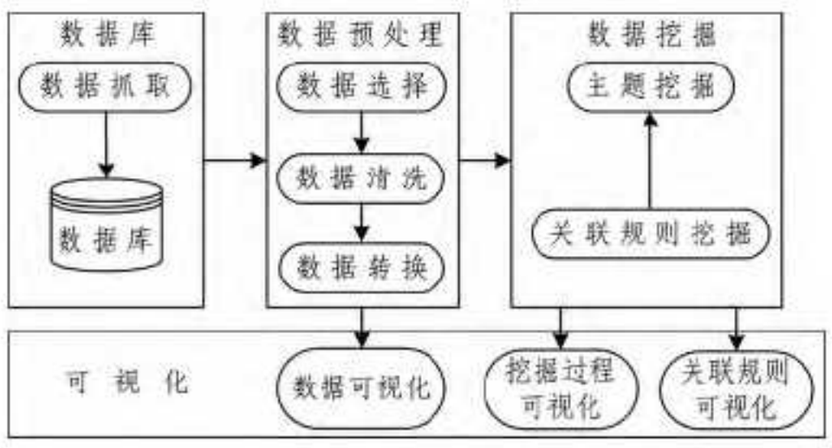

图 2 关联分析的网络数据可视化技术实现框架

关联分析的网络数据可视化技术实现框架分为以下几个阶段。

(1)数据抓取：通过网络爬虫技术从多个网络论坛上抓取数据，并将其保存到本地的数据库。

(2)数据预处理：

1)数据选择：选出与本次数据分析相关的数据；

2)数据清洗：对选择出的数据进行数据清洗，将数据转变成“干净”的数据；

3)数据转换：将清洗后的数据转换成关联规则算法所需要的格式。

(3)关联规则挖掘：使用合适的关联规则算法，对网络论

坛数据进行关联规则分析，挖掘出论坛数据集中隐藏的有价值的信息和规律。

(4)主题挖掘：使用合适的主题挖掘技术对规则下的数据进行主题挖掘，快速地发现数据集中主要的观点和话题。

(5)可视化：以适当的可视化技术，对数据集进行可视化；对关联规则结果可视化，使关联分析的结果直观易懂；通过合适的布局技术产生美观的图谱，让用户能够对可视化结果作出解释和评估，从而形成有价值的知识。

3 关联分析的网络数据可视化技术

3.1 网络论坛数据可视化

随着人们对数据可视化研究的深入，越来越多的可视化方法出现在人们的视线中。可视化技术根据可视化原理不同可以分为以下几类：基于几何的技术、面向像素的技术、基于图标的技术、基于层次的技术、基于图形的技术和基于降维映射的技术。表1从不同角度对可视化技术进行了比较。

表1 可视化技术比较

可视化技术	适用对象	可视的数据量	大数据量交叠问题	特点
基于几何的技术	数据量不大，但维数较多的数据集	较多	存在	比较容易观察数据的分布并发现其中的歧义点
基于像素的技术	大型的数据集	多	可避免	对输入的查询数据可以给出更丰富的信息，便于用户从中发现隐含的关系
基于图标的技术	维数不多，但某些维含有特别含义的数据	较少	存在	用户可以根据图标的显示更准确地理解这些维的意义
基于层次的技术	具有层次关系的多维信息和数据	适中	可避免	可以将多维空间划分为若干子空间，再对这些子空间以层次结构的方式组织并以图形表示出来
基于图形的技术	结构关系较强的数据	较多	可避免	用整个图形可以表达多维信息及其相互关系
基于降维映射的技术	数据量不大，但维数较多的数据集	较多	存在	能够可视化高维数据集并可以在低维可视空间中展现数据的整体结构和分布

国内外学者已对可视化做了很多的研究与实现。Prefuse toolkit[5]是一个可以帮助使用 java 语言的开发者开发交互的信息可视化程序；文献[6]针对力导引布局算法产生的图像经常是节点聚集在屏幕中央，难以分辨结构信息的问题，提出了子群分析布局 SAL(subgroup analysis layout)算法；文献[7]提出基于扫描线种子填充的像素可视化方法和运用四叉树管理技术对海量数据进行像素排列；文献[8]采用分层次捆绑和渲染技术在一定程度上解决了海量数据集显示时产生众多交叉重叠的问题；许彦如等人[9]提出了海量论坛数据的层次可视化，并针对网络论坛数据进行考虑偏好的网络数据可视化分析，但由于网络论坛数据的复杂性和大规模性，可视化的时间复杂度较高，界面出现混乱重叠现象，而且只是对数据进行可视化，并没有进行深入挖掘以发现其隐藏的有价值的信息。

可以看出，网络论坛数据可视化目前普遍存在界面显示混乱和系统执行时间较长的问题。为了解决上述问题，同时又考虑到网络论坛数据具有层次性的特性，本文将采用 D3 技术实现网络论坛数据的层次可视化。

3.2 关联分析网络数据

关联规则可以从数据集中发现属性间存在的、隐藏的、新颖的、有趣的关联或相关关系，进而从海量数据中获取信息和知识。

目前已有很多关联规则算法及其改进算法，比如，文献[10]提出一种借助内存减少访问数据库时间与利用辅助表减少扫描次数的改进算法，其但当数据量较大时，辅助表所占用的内存空间也相应增大。针对这个问题，文献[11]提出基于 FP-Tree 的最大频繁项集挖掘算法——NCFP-Max 算法，其通过引入单向有序的 FP-Tree 和项目表格提高了算法的时间效率和空间效率。

本文采用挖掘效率较高的基于 FP-Tree 的最大频繁项集挖掘算法，并将对其做进一步的优化。数据挖掘技术虽然能够从数据中挖掘出有价值的信息，但是挖掘过程不可见，算法对用户来说就好像是一个“黑箱子”，不利于人们对算法的理解；挖掘结果是一些数据或者规则，非技术人员很难去理解；而且，用户没有参与到数据挖掘的过程中，不能很好地实现人对挖掘结果的控制与调整。

3.3 关联规则与可视化技术的结合

在一般的可视化系统中，内容是未知的，用户是一个分析研究者，而系统将数据以可视化的形式表现出来，协助用户获得观察的结果。如果将数据挖掘和可视化技术相结合，形成可视化数据挖掘系统，则有利于人们从海量数据中提取信息。

将可视化技术应用到关联规则挖掘结果的表示中，是关联规则挖掘研究的一个新进展。近年来，人们已经提出了很多种可视化技术来支持用户对关联规则的观察和分析。这些关联规则可视化技术主要有基于表的可视化技术、基于二维矩阵的可视化技术、基于有向图的可视化技术、基于平行坐标的可视化技术、基于规则多边形的可视化技术等等。

目前，国内外学者对关联规则可视化技术不断地改进和创新。文献[12]通过以平行坐标的每条坐标轴表示对数据库的 1 次扫描，对平行坐标方法进行了改进；郭晓波[13]等提出基于概念格的多值属性关联规则可视化；文献[14]将删除冗余规则的算法与基于修补项的 FP-Growth 算法结合起来，避免了产生候选集和重复扫描数据库，但是，该算法当数据量较大时，规则间的遮蔽现象较为严重。

针对目前关联规则可视化技术的不足，文献[15]提出了基于三维坐标的关联规则可视化新技术 AR3DV，它用 Java 3D 三维坐标系中的 X 轴表示关联规则，Z 轴表示规则的项，Y 轴表示规则的支持度和置信度，并用 Java3D 的编程思想和编程步骤实现了这种表示。此技术能有效地显示大量多对多和多维的关联规则，且不会出现界面紊乱、表示歧义等问题。但它也有不足之处，即为了考虑挖掘算法的执行效率，挖掘的数据库是按某一固定的格式存储的数据集，而且系统中只有 AR3DV 一种可视化技术，不利于结果的比较和分析。关联规则的挖掘算法取自于已经成熟的算法，这些算法或多或少

存在着某些缺陷。

本文采用三维坐标的关联规则可视化技术 AR3DV，但是针对它存在的不足：数据不贴近真实数据，可视化技术单一且关联规则算法陈旧，本文将采用真实的网络论坛数据，以保证数据的真实可信，系统中存在数据可视化技术和关联规则可视化技术等多种可视化技术，还对 FP-Growth 关联规则算法进行优化，以保证算法的准确性和快速性。

3.4 设计关联规则和数据之间的交互操作

为了能让用户很好地对感兴趣的规则下的数据进行了解和观察，本文设计一种关联规则和其对应数据的交互。用户对某一规则或者某几个规则感兴趣时，通过下钻找到其对应的数据集，了解数据的详细信息；而且，可以通过上卷返回到关联规则界面。

3.5 主题挖掘

主体模型的思想源于 LSI(Latent Semantic Indexing，隐性语义索引)。在 LSI 的基础上，Hofmann 将 LSI 进化为概率模型，提出了 PLSI(Probabilistic Latent Semantic Indexing，概率隐性语义索引)，该模型被看成一个真正意义上的主题模型。而 Blei 等人在 PLSI 的基础上进行扩展，提出了 LDA(Latent Dirichlet Allocation，潜在狄利克雷分布)。LDA 是一个比 PLSI 更为完全的概率生成模型。

文献[16]提出基于 LDA 模型的 BBS 话题演化。针对传统的方法更多的是检测和跟踪话题，而没有考虑话题的演化，它提出了基于 LDA 模型的话题演化方法，表示话题在时间上的演化情况，发现热门话题和冷门话题，能更好地指导网民了解正在发生的事情。但它只考虑帖子的发布时间和内容属性，没有考虑帖子的作者、回复、链接等元数据特征，无法从其它特征中获取隐藏的话题信息。此外，文献[17]提出一种通用的框架，以发现网络论坛上的高质量主题。

本文采用 LDA 主题模型，分别对帖子的内容和帖子的作者等多个数据特征进行主题挖掘，发现规则下数据的主题，使论坛管理者和相关分析人员可以快速地发现感兴趣或是认为有价值的话题。

4 目前存在的问题与挑战

4.1 存在的问题

目前已有很多优秀的可视化工具，例如文献[18]在 Java 平台上开发了一个可视化数据挖掘工具 DBMiner；谷歌趋势(Google Trends)是揭示数据关联趋势的可视化新服务[19]；数据挖掘系统 SGI 公司的 MineSet 挖掘器和 IBM 公司的 Intelligent Miner 挖掘器[20]将关联规则与数据可视化捆绑在一起，既能对结果进行评估，又能实时调整挖掘过程中的算法及其参数。但是这些工具依然存在一些亟待解决的问题。在数据可视化中大致存在以下几个核心问题：

1. 数据显示时，界面上会出现大量的交叉重叠现象。采样技术、可视化的布局技术等技术手段可以在一定程度上解决这个问题。

2. 很多数据挖掘算法的时间复杂度较高，严重影响算法的运行效率。

3. 对数据进行单一的数据挖掘或是可视化展示，已无法满足当今数据分析日益增长的需求。

4. 缺少针对专一领域的可视化分析工具来促进该领域更好地向前发展。

4.2 面临的挑战

根据信息可视化的十大原则，可视化技术与应用还应该继续向以下 4 个方面努力，即直观化：直观、形象地呈现数据；关联化：挖掘、突出呈现数据之间的关联；艺术化：增强数据呈现的艺术效果，符合审美规则；交互化：增强人机交互，实现即时数据操作。由此观之，数据可视化面临的挑战大致包括以下 3 点：

1. 海量高维数据集的可视化。大量复杂的网络数据无论是在关联分析还是数据可视化方面都是一个巨大的挑战，常常出现数据处理速度慢和界面混乱重叠现象。因此，我们迫切需要找到更加高效的数据挖掘和可视化布局算法。

2. 实时分析可视化。越来越多的实时数据在网上产生，或由穿戴设备产生。如何处理并合理利用这些数据显然是另一大挑战。

3. 预测分析。预测模型的需求在与日俱增，但支持预测的系统却寥寥无几。可见，先进、可定制的可视化和实时、预测分析是未来的发展趋势。

结束语 当今异构的、非结构化的混杂网络数据的出现，对可视化提出了一个大挑战。因此，本文试图将可视化技术和关联规则挖掘技术结合，提出基于关联分析的网络数据可视化技术实现框架，合理运用数据挖掘的公式和算法，并对数据分析的过程及结果进行可视化展现，再通过一定的交互操作与主题挖掘，构建一个直观的、交互性强的、能为用户提供决策支持的可视化系统。此外，对框架中相关技术及其发展现状进行了分析与综述，并且探讨了现阶段的数据可视化工作面临的问题与挑战。

参 考 文 献

[1] 陈为，沈则潜，陶煜波. 数据可视化[M]. 北京：电子工业出版社，2013

[2] 曾悠. 大数据时代背景下的数据可视化概念研究[D]. 杭州：浙江大学，2014

[3] 郝竞超，王朝坤，司德谭，等. MMDVis：一个基于微博用户的多博文传播分析及可视化系统[J]. 计算机研究与发展，2013，50(suppl.)：399-404

[4] 余韬，肖仰华，徐晓，等. Graph Explorer：基于结构的大型网络可视系统[J]. 计算机研究与发展，2011，48(S3)：421-424

[5] Heer J，Card S K，Landay J. Prefuse：a toolkit for interactive information visualization[C]// ACM Human Factors in Computing Systems CHI 2005. 2005

[6] Wu P，Li S K. Layout algorithm suitable for structural analysis and visualization of social network[J]. Journal of Software，2011，22(10)：2467-2475

[7] 杜珊珊. 基于像素的海量数据可视化服务[D]. 秦皇岛：燕山大学，2012

[8] 刘大海. 海量数据可视化方法的研究[D]. 天津：天津大学，2009

[9] 许彦如. 考虑偏好的网络数据可视化分析[D]. 上海：华东师范大学，2012

[10] 乌文波. 应用 Apriori 关联规则算法的数据挖掘技术挖掘电子商务潜在客户[D]. 杭州：浙江工业大学，2012

[11] 王芳. 基于 FP-Tree 的最大频繁项集挖掘算法研究[D]. 南宁：广西大学，2013

[12] 钟志文. 基于平行坐标的关联规则挖掘技术可视化研究与实现[J]. 常州工学院学报,2012,25(2):29-33
[13] 郭晓波,赵书良,赵娇娇,等. 基于概念格的多值属性关联规则可视化[J]. 计算机应用,2013(8):2198-2203
[14] 吴天真. 基于修补项的关联规则可视化挖掘方法的研究[D]. 武汉:华中师范大学,2009
[15] 易先卉. 关联规则可视化技术的研究及实现[D]. 长沙:湖南大学,2008
[16] 石大文. 基于 LDA 模型的 BBS 话题演化[J]. 工业控制计算机,2012,25(5):82-84
[17] Chen Y, Cheng X Q, Yang S. Finding high quality threads in Web forums[J]. Journal of Software,2011,22(8):1785-1804
[18] 孟海东,林志举,徐贯东. 可视化数据挖掘工具的设计与实现[J]. 计算机与现代化,2011(6):132-135
[19] 张浩,郭灿. 数据可视化技术应用趋势与分类研究[J]. 软件导刊,2012,11(5):169-172
[20] 王华金,蔡虔. 数据挖掘可视化技术综述[J]. 科技广场,2009(1):235-237

(上接第 464 页)

程(一系列的重复字符)变长。

$$\text{压缩率}=\frac{\text{使用本文方法得到的数据大小}}{\text{使用传统方法得到的数据大小}}\times 100\% \quad (16)$$

②索引大小。为了获得数据 SFP 的顺序,本文方法通过垂直编码对 ψ 进行存储。表 2 列出 SA 和 ψ 的大小。SA 的数据 ψ 的大小会随着数据库中歌曲数量的增长而增长,也就是说,随着数据库的增大,压缩效率会变低。如果因为提高音乐的数量而增加了子乐纹的种类,那么相邻的数据也不一定是相同的,即便它存在于排序好的数据中。换句话说,单调增加的部分减少了,这导致了压缩效率的恶化。

表 2 索引大小

歌曲(首)	传统方法(MB)	本文方法(MB)	压缩率(%)
2000	57.3	46.8	81.68
4000	123.3	103.6	84.02
8000	254.1	217.5	85.60
12000	387.0	336.1	86.85

③总数据大小。总数据的大小如表 3 所列。总体来说,对于数据库中的每首歌,压缩率在 60% 左右。随着歌曲数量的增加,压缩率也略有提升,这归功于子乐纹压缩率的高度。

表 3 总数据大小

歌曲(首)	传统方法(MB)	本文方法(MB)	压缩率(%)
2000	115.7	70.8	61.19
4000	248.8	150.3	60.41
8000	512.4	305.0	59.52
12000	778.2	458.7	58.94

3)搜索时间

在检索时,用于查询的歌曲和数据库的是一样的,并且和原始歌曲有相同的长度。在本节中,使用每 10 秒的查询结果。也就是说,对于一个有 S_q 秒的查询乐曲,如果搜索总共花费了 S_s 秒,那么每 10 秒的查询时间可以表示为 $S_s/S_q\times 10$。

表 4 列出了所有音乐数据的集合中每首歌曲的平均搜索时间。减速因子(SLF)指的是采用本文方法所消耗的时间与传统方法相比的倍数,即 SLF 越小,本文方法越快。SLF 的计算如式(17)所示。

$$SLF=\frac{\text{使用本文方法的搜索时间}}{\text{使用传统方法的搜索时间}} \quad (17)$$

表 4 平均搜索时间

歌曲(首)	传统方法(ms)	本文方法(ms)	SLF
2000	1.1	12.9	11.7
4000	2.3	18.5	8.0
8000	3.8	23.1	6.1
12000	5.0	25.5	5.1

表 4 还表明本文方法搜索会花费更多时间,这主要是数据结构的关系。然而,随着数据库中歌曲的增多,减速因子呈下降趋势。

结束语 本文提出一种基于游程编码和垂直编码来压缩乐纹后缀数组的索引压缩方法。实验结果表明,该方法能够有效地节省大量的乐纹数据库空间。尽管该方法在检索时需要花费更多的查询时间,但查询时间的倍数会随着数据库中歌曲数量的增加而呈下降趋势。因此,我们下一步工作将包括在海量数据中如何提高检索速度以及数据库音乐检索的具体应用研究等。

参 考 文 献

[1] Casey M A, Veltkamp R, Goto M, et al. Content-based music information retrieval: current directions and future challenges[J]. Proceedings of the IEEE, 2008, 96(4): 668-696
[2] Xiao Q, Xin Luo, Saito N, et al. Index Compression for Audio Fingerprinting Systems Based on Compressed Suffix Array [J]. International Journal of Information and Education Technology, 2013, 3(4): 455-460
[3] Bellettini C, Mazzin G. A Framework for Robust Audio Fingerprinting [J]. Journal of Communications, 2010, 5(5): 409-424
[4] Xiao Qing-mei, Saito N, Luo Xin, et al. Index Compression for Audio Fingerprinting Systems Based on Compressed Suffix Array [J]. International Journal of Information and Education Technology, 2013, 3(4): 455-460
[5] Haitsma J, Kalker T. Ahighly robust audio fingerprinting system [C]// Proc. 3rd International Conference on Music Information Retrieval. 2002
[6] 李伟,李晓强,陈芳,等. 数字音频乐纹技术综述[J]. 小型微型计算机系统,2008,29(11):2124-2130
[7] Wang A L. An Industrial-Strength Audio Search Algorithm [C]// Proc. the 4th International Conference on Music Information Retrieval(ISMIR 2003). 2003
[8] Miller M, Rodriguez M, Cox I. Audio fingerprinting: Nearest neighbor search in high dimensional binary spaces [J]. Journal of VLSI Signal Processing, 2005, 41(3): 285-291
[9] Manber U, Myers G. Suffix arrays: a new method for on-line string searches [C] // 1st ACM-SIAM Symposium on Discrete Algorithms. 1990
[10] Jensen R, Shen Q. Semantics-preserving Dimensionality Reduction: Rough and Fuzzy-rough-based Approaches [J]. IEEE Transactions on Knowledge and Data Engineering, 2004, 16(12): 1457-1471
[11] 徐英进,蔡锐,蔡莲红. 一种基于“乐纹”的海量音乐检索系统[C]//第二届和谐人机环境联合学术会议(HHME2006)——第 15 届中国多媒体学术会议(NCMT'06)论文集. 北京:清华大学出版社,2006
[12] 姚全珠,张楠,杨增辉,等. 基于压缩后缀数组技术的搜索引擎[J]. 计算机工程,2008,34(10):83-85