学校编码：10384　　分类号＿＿＿密级＿＿＿
学号：X200343052　　UDC ＿＿＿

厦门大学

硕 士 学 位 论 文

基于关联规则的社会性网络行为研究

Research on Social Network Behavior
Based on Association Rules

俞燕燕

指导教师姓名：李 绍 滋 教授
专 业 名 称：计 算 机 应 用
论文提交日期：2006年 12月
论文答辩时间：2007年 月
学位授予日期：2007 年 月

答辩委员会主席：＿＿＿＿＿＿
评 阅 人：＿＿＿＿＿＿

2007 年 月

厦门大学学位论文原创性声明

兹呈交的学位论文，是本人在导师指导下独立完成的研究成果。本人在论文写作中参考的其他个人或集体的研究成果，均在文中以明确方式标明。本人依法享有和承担由此论文产生的权利和责任。

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人完全了解厦门大学有关保留、使用学位论文的规定。厦门大学有权保留并向国家主管部门或其指定机构送交论文的纸质版和电子版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆被查阅，有权将学位论文的内容编入有关数据库进行检索，有权将学位论文的标题和摘要汇编出版。保密的学位论文在解密后适用本规定。

本学位论文属于

1、保密（ ），在　　年解密后适用本授权书。

2、不保密（ ）

（请在以上相应括号内打“ √ ”）

作者签名：　　日期：　　年　月　日

导师签名：　　日期：　　年　月　日

摘　　要

随着Web和Internet的迅猛发展，网络规模不断扩大，信息爆炸式增长，人们对资讯需求的深化，用户需求的细分将成为未来发展的趋势。如何解决信息过载；同一类别用户的行为（即社会性行为）存在着特定的关系和规则的，如何发现关系和规则。根据用户的不同的文化背景、教育背景、地域背景提供合适其兴趣的信息将变得非常有价值。人们希望能够提供更高层次的数据分析功能，自动和智能地将待处理数据转化为有用的信息和知识，数据挖掘与知识发现就为迎合这种需求而产生并迅速发展起来的。

关联规则挖掘是数据挖掘的重要分支。自从Agrawal等学者于1993年首先提出了关联规则挖掘问题以来，诸多研究人员对关联规则挖掘问题进行了大量的研究，提出了很多高效的算法。

本文在现有的关联规则挖掘研究的基础上作了如下工作：

1、在查阅国内外大量文献资料的基础上，对数据挖掘技术和关联规则挖掘技术进行了概述，讨论了关联规则挖掘的典型算法Apriori算法，以及频繁项集算法的几种优化方法的基本思想和执行情况。

2、关联规则算法的改进。本文提出了一种基于散列的快速AprioriTid改进算法（AprioriTid_Hash），该算法在AprioriTid算法的基础上，采用基于候选项L_k的地址的哈希映射方法，从而提高了算法的执行效率。通过与AprioriTid算法做了性能比较，实验验证该算法是有效的。

3、关联规则技术与社会性网络行为相结合。以一个音乐共享平台来分析用户在线收听歌曲的社会性网络行为，对模型MusicMiner实现细节进行了描述，从数据挖掘的角度出发，对原始数据进行预处理，将数据仓库和数据挖掘技术相结合，应用关联规则的AprioriTid_Hash算法对大量的网络行为数据进行挖掘，得到了一些有用的关联规则。

关键字：数据挖掘；社会性网络行为；关联规则

Abstract

With the rapid development of Web and Internet, enlarged network scale and increased information explosively, deepen information requirement and the sub-user demand will become the future development trend. How to solve information over loading; a kind of user's behavior (namely sociality network behavior) exist relation and rules specifically, how to discover relation and rules. It is worth providing information of one's interest based on user's different background of the civilization, education and clime. So it's expected to provide the function of data analysis even with more high administrative levels, which can change processing data into useful information and knowledge automatically and intelligently. Data mining and knowledge discovery in database (KDD) get rapid development for their cater to this demand.

Association rule mining is an important branch of the data mining. Since association rule mining was proposed by Agrawal and other scholars, it has been conducted a lot of research and many researchers put a lot of efficient algorithms in data mining.

Based on the existing mining association rules, the following work is done:

1. The background knowledge of data mining and association rules mining is introduced briefly; State the typical algorithm of association rules mining --Apriori algorithm; and discuss the basic idea and the implementation of several frequent item-set optimize methods.

2. The improved algorithm for mining association rules. This paper presents a fast AprioriTid (AprioriTid_Hash) , which is based on the hash algorithm, enhanced the efficiency of the algorithm using candidate set L_k address-mapping approach. Comparing with the performance of AprioriTid algorithm, the experiments show that the method is effective.

3. Combining association rules technique with sociality network behavior, analyse the social network behavior of users to listen to songs on line with a platform to share music and describe the achievement in details of the model Music Miner. From the perspective of data mining, the raw data is pretreated and data warehouse is connected to data mining. Applying association rules AprioriTid algorithm, a lot of network behavior data is mined and some useful association rules are generalize.

Keywords: data mining; sociality network behavior; association rules

目 录

Contents

第一章　引言

1.1 研究背景

近年来以TCP/IP为核心的Internet取得了飞速的发展，网络的规模不断扩大、网络新业务不断出现，给人类社会带来了巨大的变化和影响。Internet已经从原来的科研用途进入了寻常普通大众生活，许多原来的社会活动开始在Internet上展开，既有商务行为，如网上书店、电子拍卖、网络旅游服务等等；也有文化娱乐，如网络影院、网络音乐。

在网络规模不断扩大，信息爆炸式增长的情况下，如何解决信息过载，根据用户的不同的文化背景、教育背景、地域背景提供合适其兴趣的信息将变得非常有价值。

目前的搜索引擎主要基于关键字为用户提供信息，但是在很多时候，用户并不能准确的提供合适的关键字来获取信息。

同一类别用户的行为（即社会性行为）存在着特定的关系和规则，以前要对同一类别用户的已经发生的行为来进行分析由于数据的采集成本及技术限制是非常困难的，现在大量的社会活动开始在Internet上展开，数据采集成本已经大大降低，同时信息爆炸式增长，已经具备了对社会性行为进行分析的可行性和必要性。

数据挖掘与知识发现就是在这样的应用背景下产生并迅速发展起来，提供更高层次的数据分析功能，自动和智能地将待处理数据转化为有用的信息和知识，以达到分析已有数据发现内在关系和规则，并用来辅助用户发现新的信息，起到海量信息的过滤、提高效率的目的。

1.1.1 数据挖掘的含义

数据挖掘是从大量数据中挖掘出隐含的、先前未知的、对决策有潜在价值的知识和规则的高级处理过程。通过数据挖掘，有价值的知识、规则或高层次的信息就能从数据库的相关数据集合中抽取出来，并从不同角度显示，从而使大型数据库作为一个丰富、可靠的资源为知识的提取服务。

1.1.2 社会性网络行为分析的意义

网络行为[1]，即网络空间主体的行为，包括交易行为、消费行为、娱乐行为、

政治行为、违法行为等。由于这些行为又是编码（技术）所控制和允许的行为，因此，它们也成为了信息科学与技术学科的研究对象。

社会性网络行为分析[2]指：对基于互联网开展的社会性行为进行用户行为信息的采集，并对收集的用户网络行为信息通过关联规则挖掘算法、类聚算法给予分析，最大程度地挖掘信息的价值。

在网上，由于有几乎无穷多的产品可以选择，通过对社会性网络行为的分析，用户真正喜好的产品能被挖掘出来。

首先我们来看一个案例[3]。1988 年，英国登山家Joe Simpson 写了一本名叫《触摸巅峰》（该书讲述了在秘鲁安第斯山脉发生的一次与死神擦肩而过的登山事故。）这本书颇受好评， 但不太畅销， 并很快就被人们淡忘了。可是十年后， 有趣的事发生了。Jon Krakauer 写的另一部描写登山悲剧的书《进入稀薄空气》成为了畅销书。突然间读者又开始对《触摸巅峰》产生了兴趣。

Amazon (亚马逊)的推荐造成了这个现象。网上书商Amazon 的软件注意到不少喜欢《进入稀薄空气》的读者也喜欢《触摸巅峰》，就向购买《进入稀薄空气》的所有读者推荐《触摸巅峰》。读者接受了推荐，并且真的觉得这本书很好。一时间， 网站的留言里好评如潮， 销量因此进一步增加， 这就带来了更多的好评， 于是形成了一个正反馈。

这一案例不仅仅适用于网上书商，它其实揭示了一个全新适用于媒体和娱乐业的经济模型。以流行为主导的经济是旧时代的产物， 在这个时代里， 没有足够的库存空间来存放所有的东西， 以满足每个人的需要。比如说， 没有足够的货架来存放所制作的CD、DVD 和游戏；没有足够的银幕来放映所有的电影；没有足够的频道来播放所有的电视节目； 也没有足够的无线电波段来播放所有的音乐； 甚至没有足够的时间把所有的内容传播给用户。

这是一个资源稀缺的世界。但是， 随着在线分销和零售模式的出现， 我们正迈入一个资源极大丰富的世界。

但事实上， 我们对于自己想要什么也并不是很清楚。为了了解我们那种不受资源稀缺的经济所限制的真正口味，推荐人们真正喜欢的产品，所以社会性网络行为的分析是非常有意义的。由于基于网络和计算机，因此社会性网络行为分析具有采集及时丰富、信息传输高效和可操作性强等特点，具有较高的应用价值。

至于采用怎样的关联规则挖掘算法就更为重要了。

1.1.3 关联规则挖掘

关联规则挖掘的任务是在事务数据库D中找出满足用户给定的最小支持度minsup和最小置信度minconf以及用户感兴趣的、有用的关联规则。是面向特定领域，特定前提、约束条件的规则，同时还要能够易于被用户理解，并能用自然语言表达所发现的结果。

1.2 挖掘关联规则的难点

挖掘关联规则时主要要解决下面两个问题。

首先是算法的复杂性。目前的挖掘关联规则的算法都是针对这个问题而提出来的。通常提出的算法从两个方面来考虑如何提高算法的效率。（1）减少I/O操作。关联规则挖掘GB甚至TB数量级，频繁的I/O操作势必会影响关联规则的挖掘效率。减少扫描数据库D的次数可以减少I/O操作，提高效率；（2）降低需要计算支持度的候选项目集的数量，使其与频繁项目集的数量接近。候选项目集数量的减少可以节省为处理部分候选项目集所需要的计算时间和存储空间。

其次是如何从产生的规则中选择用户感兴趣、有用的规则。最小置信度和最小支持度并不能确保所挖掘出来的关联规则都是用户感兴趣的，其中可能包含许多冗余、无意义的关联规则。而且支持度和置信度较高的关联规则有可能是常识的知识，不能称之为信息。因此制定好的关联规则兴趣度计算标准可以使挖掘出的关联规则更能满足用户的需求。

1.3 本文的工作

数据挖掘技术从一开始就是面向应用的。关联规则挖掘在商业等领域的成功应用，使它成为数据挖掘中最重要、最活跃的研究内容。国内外在关联规则挖掘方面的研究已经取得了较大的进步，但关联规则挖掘技术在具体领域中的应用还存在着不足，如内存问题，需要进一步研究和提出更好的解决方案。

本文从挖掘关联规则的核心问题出发，以研究社会性网络行为为目的，进行了较为深入的研究。

以一个音乐共享平台来分析用户在线收听歌曲的社会性网络行为，参考模型MusicMiner。在已架构的音乐共享平台，用户搜索，播放收听歌曲，系统记录保

存所有参与该平台的用户在线收听歌曲的相关信息，根据用户（UserId）建立一个相应的知识库。将关联规则挖掘技术和领域背景知识应用在发掘用户网络行为中,对用户（UserId）在线收听歌曲记录来进行分析，同样先找出高频项目组再经由判断最小信赖度看其关联规则是否成立，如果成立，系统就利用这些关联规则给用户推荐其可能喜欢的相关歌曲。

本文重点研究关联规则及其相关算法，主要完成了以下几方面的工作：

1．介绍了关联规则及其相关概念，讨论了目前较有影响的Apriori算法以及频繁项集算法的几种优化方法的基本思想和执行情况。如AprioriTid算法、Hash方法。

2．关联规则算法的改进，本文提出了一种基于散列方法的AprioriTid算法（AprioriTid_Hash），能较为快捷的求解频繁项目集，由于分别依次对C_{k-1}每个事务集内的L_{k-1}进行连接join操作，大大减少了n的值，降低了时间复杂度$O(n^2)$，同时在散列时就能得到频繁k-项集集合L_k，省去了找出其支持度不小于最小支持度的频繁项目集L_k所要花费的时间。并与AprioriTid进行性能比较，实验证明了AprioriTid_Hash算法是有效的。

3．应用AprioriTid_Hash算法，分析社会性网络行为。本文基于音乐共享平台中网络行为的大量数据，从数据挖掘的角度出发，对原始数据进行预处理，以数据仓库和数据挖掘技术相结合，对模型MusicMiner实现细节进行描述。应用关联规则的AprioriTid_Hash算法对大量数据进行挖掘，并得到了一些有用关联规则。

本文的创新点如下：

1．提出了基于散列的关联规则ApripriTid改进算法，相对于ApripriTid算法有了很大的性能提高。

2．关联规则与社会性网络行为相结合，对大量的网络行为数据进行收集、整理、描述、显示和分析统计，采用关联规则算法的核心思想，对每个用户访问行为、频度、内容等的分析，探索数据的内在数量规律性。

同时也为将来对社会性网络行为研究作了一些思维方式上的探讨。通过对社会性网络行为进行分析，总结出用户习惯，并根据用户习惯、特征，分析用户之

间的共性给用户提供更多的相关信息资源，让用户能获得更多的有益的信息来源，具有现实意义。

1.4　本文的组织结构

本文的工作围绕关联规则及应用关联规则分析网络行为展开，全文共分为五章。论文结构安排如下：

第一章引言，简单介绍了社会性网络行为分析研究背景，关联规则研究状况，以及本文的研究思路和全文的组织。

第二章数据挖掘概述，介绍了数据挖掘背景、数据挖掘的概念、功能及分类，目前数据挖掘的研究方向。

第三章介绍了关联规则及其相关算法，分别介绍了关联规则的提出；关联规则的概念、种类、关联规则价值衡量的方法；关联规则生成及算法思路；重点介绍了频繁项目集算法及其相关算法。

第四章应用关联规则分析网络行为，也是将关联规则技术面向应用。主要是模型MusicMiner实现细节的描述，从数据挖掘的角度出发，对原始数据进行预处理，将数据仓库和数据挖掘技术相结合，应用关联规则的AprioriTid_Hash算法对大量数据进行挖掘，得到了一些有用关联规则。

第五章结束语，指出了进一步研究的方向。

最后致谢、参考文献和附录。

第二章 数据挖掘概述

2.1数据挖掘背景

2.1.1 知识发现出现背景

数据挖掘是20世纪90年代兴起的一项新技术，它是知识发现过程的关键步骤。知识发现的提出，有着很深刻的社会背景。

随着信息技术发展以及社会的不断发展，在支配人类社会三大要素(能源、材料和信息)中，信息愈来愈显示出其重要性和支配力，它将人类社会由工业化时代推向了信息化时代，使现代社会所有大的机构都卷入到数据处理(数据搜集、存储、检索、传送、分析和表示)的浪潮中。同时随着人类活动范围扩展，节奏加快，数据和信息量以指数形式向上增长。特别是80年代以来，先进可靠的数据库技术出现，越来越多的数据通过计算机存储，人们能以廉价的方式快速方便的获取和存储数据。九十年代互联网(Internet)的出现和发展，以及随之而来的企业内部网(Intranet)和企业外部网(Extranet)以及虚拟私有网(VPN--Virtual Private network)的产生和应用，将整个世界联成一个小小的地球村，人们可以跨越时空地在网上交换信息和协同工作。这样，展现在人们面前的已不仅仅是局限于本部门、本单位和本行业的庞大数据库，而是浩瀚无垠的信息海洋。面对这极度膨胀的数据信息量，人们感到“数据过剩”、“信息贫乏”。因此，人们迫切地希望找到一些从大量数据中发现知识的方法和技术。知识发现就在这样的社会背景应运而生。

2.1.2 知识发现和数据挖掘

如何对数据与信息快速有效地进行分析、加工、提炼，以获取所需知识并发挥其作用，向计算机和信息技术领域提出了新的挑战。

其实计算机和信息技术发展的过程，也是数据和信息加工手段不断更新和改善的过程。早年受技术条件限制，一般用人工方法进行统计分析，和用批处理程序进行汇总和提出报告。

数据仓库的出现，为更深入对数据进行分析提供了条件，针对市场变化的加速，人们提出了能实时分析和产生报表的在线分析手段OLAP(On Line Analytical Processing)，它是一种友好而灵活的工具，它能允许用户以交互方式浏览数据