多关系频繁模式发现研究*)

张 伟 杨炳儒 钱 榕
(北京科技大学信息工程学院 北京 100083)

摘 要 频繁模式发现是数据挖掘的重要任务之一。现实数据通常存储于由多个关系组成的关系数据库中。传统的频繁模式发现方法只能直接完成单一关系中的模式发现,如果要完成多关系数据的挖掘,会产生操作复杂性和信息丢失等问题。多关系数据挖掘是当前数据挖掘研究中快速发展的重要领域之一。多关系频繁模式发现方法能够直接从复杂结构化数据中发现涉及多个关系的复杂频繁模式,避免了传统方法的局限。本文首先归纳多关系频繁模式发现方法的发生历史背景,其次分析总结多关系频繁模式发现方法,最后提出了多关系频繁模式发现将来发展需重点解决的问题和面临的挑战。

关键词 多关系数据挖掘,频繁模式发现,归纳逻辑程序设计,选择图,基于图的数据挖掘

A Survey of Multi-relational Frequent Pattern Discovery

ZHANG Wei YANG Bing-Ru QIAN Rong
(School of Information Engineering, Beijing University of Science and Technology, Beijing 100083)

Abstract Frequent pattern discovery is one of the most important tasks of data mining. Most of today's structured data is stored in relational databases including multiple relations. Traditional approaches look for frequent patterns in a single relation, and it is usually difficult to convert multiple relations into a single relation without losing essential information. Multi-relational data mining is one of rapidly developing subfields of data mining. Multi-relational frequent pattern discovery approaches directly look for frequent patterns that involve multiple relations from a relational database. This paper provides a survey of the research of multi-relational frequent pattern discovery. Firstly, the background and context that multi-relational frequent pattern discovery arises from is analyzed. Secondly, typical algorithms of multi-relational frequent pattern discovery are introduced and analyzed. Finally, several challenging researching problems are identified.

Keywords Multi-relational data mining, Multi-relational frequent pattern discovery, Inductive logic programming, Selection graph, Graph-based data mining

1 引言

数据挖掘作为知识发现的核心步骤,致力于发现海量数据中隐藏的模式。频繁模式发现是数据挖掘的重要任务之一,早期的相关研究成果包括 Apriori 算法[1]及其变体[2]。这类方法的知识表示方式主要是命题逻辑形式系统并且只能从单一关系中发现模式。但是,大多数现实关系数据库中的信息存储于多个关系中,当在多关系数据中发现模式时,模式自然地要涉及多个关系。若使用这类经典数据挖掘方法,应把数据先从多个关系中纳入一个单关系中,然后才能进行挖掘。这不仅需要大量的预处理工作和谨慎的设计,并且可能导致信息丢失、语义偏差以及效率降低等问题。另外,许多复杂模式难以用命题逻辑语言表示。

有别于上述经典方法,另一类频繁模式发现方法是来自多关系数据挖掘领域的多关系频繁模式发现方法。

多关系数据挖掘(Multi-relational Data Mining, MRDM)方法,发现关系数据库中涉及多个关系的复杂模式,且直接在多个关系上分析数据而无需向单一数据表的转换[3~6]。作为一个新的数据挖掘研究领域,多关系数据挖掘是一个跨学科领域,吸纳了归纳逻辑程序设计(Inductive Logic Programming, ILP)、KDD、机器学习和关系数据库的研究成果,致力于处理由多关系组成的关系数据库知识发现问题,研究挖掘多关系型数据的新型技术及其有效的应用实践。经过近几年的研究,数据挖掘中通常解决的各种任务及其使用的方法已经扩展到多关系情况下。当前的 MRDM 方法考虑到了所有主要的数据挖掘任务,包括频繁模式发现、分类、回归、聚类、统计模型学习等。

多关系频繁模式发现是一个重要的多关系数据挖掘任务,既可以作为独立的模式在各种领域得到应用,同时其发现过程也可作为完成其他多关系数据挖掘任务之方法的核心步骤,如关联分析、频繁子句发现、序列模式发现、时序模式发现、某些多关系命题化方法以及某些分类方法。本文将围绕多关系频繁模式发现方法展开论述。首先归纳总结多关系频繁模式发现方法的发生背景,其次分析典型的多关系频繁模式发现方法及其新近发展,最后提出了多关系频繁模式发现将来发展需重点解决的问题和面临的挑战。

2 多关系频繁模式发现方法产生的历史背景

不论是经典频繁模式发现方法 Apriori 算法及其变体,还是多关系频繁模式发现方法,它们的提出,实际上都是基于各自不同的数据挖掘方法论和研究范式的。系统地理解多关系频繁模式发现方法的本质特征,应该首先明确多关系数据挖掘这一领域发生发展的历史脉络及其一般方法论。

从历史的角度看,经典数据挖掘方法是基于"属性-值学

*)获《国家科技成果重点推广计划》项目(2003EC000001)资助。张 伟 博士研究生,研究方向:数据挖掘;杨炳儒 教授,博士生导师,研究方向:知识发现与智能系统、柔性建模与集成技术;钱 榕 博士研究生,研究方向:数据挖掘。

习”方法的，而多关系数据挖掘是在“关系学习”发展的背景下产生发展起来的。从“属性-值学习”向“关系学习”发展是多关系数据挖掘产生的历史背景，本节剩余部分描述了这一发展的原因，从而为理解具体的多关系频繁模式发现方法奠定了基础。

在属性-值学习情况下，每一样例以属性-值元组的形式表示。在这种形式下，属性种类是固定的，每个属性有一个给定的值相对应，从而整个数据集可以被看作为关系数据库中的一个表或关系，表中的每一行相应于一个样例，而每一列相应于一个属性。假设语言是命题逻辑语言，相应的命题以如下形式给出：“Attribute ⊕ value”，其中的⊕是预先定义的一个操作符集合{<，>，=}中的一个元素。因此，属性-值学习使用的算法又被称为命题化算法。

与上述情况相反，实际应用中的关系数据库为了有效地组织和访问数据，数据以多关系的形式组织。一个样例信息由位于多个关系中的多个元组描述。进一步来看，关系数据库中的结构表达了位于不同关系中的元组间的联系，而这种联系体现了问题域中某些重要的背景知识和样例信息的结构内容。属性-值学习的单表假定无法直接利用这类联系及其蕴涵的信息内容，因而无法发现现实世界数据中隐藏的更为复杂的模式。

虽然在原则上，多个关系表可以集成到一个单关系表中，但在实践中这一方法存在许多问题。文[7～10]从不同角度详细分析了相关问题。我们归纳如下。

一般情况下，如果使用传统命题学习算法，组成一个关系数据库的多个关系表就应集成到一个单关系表中。有如下两种方法实现这种集成：

1)在所有的关系上通过关系连接操作重构一个单一的泛关系。这一方法有许多潜在问题：

(1)计算泛关系的时空代价异常大。最终得到的泛关系中，存在大量冗余数据。更为重要的是，整体数据量与原始数据比，异常地大。最糟情况下，最终数据量随数据库中原始关系数量与单一关系中元组数呈指数生长，这加剧了海量数据处理问题的难度。

(2)在多对一关系形成的泛关系中，一个样例由多行组成，按照属性-值学习方法的每一行代表一样例的假定，结果会出现语义偏差。另外，对于涉及自连接的关系表，难以确定其连接深度。

(3)如果将样例的所有信息放入结果关系的单一元组中，则对于复杂数据库，一方面会出现大量空值属性，另一方面(更为重要)非常难以确定结果关系的全部属性。

(4)数据重复导致统计偏差。

(5)在许多种问题上，属性-值学习效率非常低。

2)为避免冗余，可以在某些表上做聚合操作，所得聚合值代表这些表中的信息加入一个核心关系表中，而这些表的原始元组不必加入。但是仍有两个问题：

(1)许多细节信息在聚合操作后丢失了；

(2)聚合属性的选择需要对问题域有良好的理解。如果理解出现错误，那么与使用细节信息的情况相比，结果会较差。

最后值得指出的是，属性-值学习使用的命题逻辑知识表示方式在表达复杂模式方面表达不简洁，表达力差。

因此，机器学习与数据挖掘技术明确地需要考虑学习任务的关系表示方式及其相关搜索机制，即直接在多关系数据集上学习涉及复杂关系结构的模式。在机器学习领域，这类学习问题及其解决方法被称为关系学习(Relational Learning)。在这领域，这类挖掘方法的研究形成了多关系数据挖掘。如同KDD的早期发展受机器学习领域的研究影响一样，多关系数据挖掘是在关系学习发展的背景下产生发展起来的。

在关系学习方法中，发展最早同时也是研究相对最成熟的领域是归纳逻辑程序设计(ILP)方法[11～13]。ILP的方法和技术在解决各种数据挖掘任务中得以进一步发展，在这一趋势下，ILP各方面的研究和拓展形成了多关系数据挖掘[3]研究领域。历史上多关系数据挖掘也被称为关系数据挖掘(Relational Data Mining，RDM)。多关系数据挖掘在其后发展中也吸纳了多种非ILP方法，形成了一个具有多方法谱系的研究领域。

3 多关系频繁模式发现ILP方法

本节介绍多关系频繁模式发现的ILP方法。

多关系频繁模式发现的最早期研究成果是算法WARMR[14,15]。WARMR于1997年由L. Dehaspe和H. Toivonen提出。该算法基于ILP技术，是经典频繁模式发现算法Apirior向多关系数据挖掘的更新，能够从多关系数据中发现一阶逻辑表示的频繁模式。后来的研究进一步使用这一方法解决图数据中发现频繁连接子图的任务[16]，并应用于毒物学领域化学分子的结构研究，取得了较好的效果。

WARMR奠定了多关系频繁模式发现ILP方法的研究基础，在其后的该领域ILP方法研究可以看作是基于WARMR的性能优化和功能扩展。3.1节介绍多关系频繁模式发现任务定义和相关ILP概念；3.2节分析WARMR核心算法；3.3节归纳分析多关系频繁模式发现ILP方法的研究进展。

3.1 多关系频繁模式发现任务定义

首先，我们给出多关系频繁模式发现任务形式化定义。以这一形式化定义为基础，我们详细描述WARMR算法的具体内容。

1997年，文[17]给出了一个数据挖掘任务的一般化形式定义。基于这一定义，结合WARMR算法与其他多关系频繁模式发现算法，我们归纳定义多关系频繁模式发现如下：

定义1 给定一个多关系数据库D、一个多关系模式表示语言L、一个选择谓词q，多关系频繁模式发现任务即为，发现一个关于D、L、q的理论$Th(L, D, q)$，使得$Th(L, D, q)=\{Q \in L \mid q(D, Q)$为真$\}$。$q(D, Q)$为真当且仅当$Q$在$D$中的频度不小于频度阈值。

在WARMR中，多关系数据库中的数据以及多关系频繁模式的表示使用DATALOG。DATALOG[18]是一阶子句逻辑(First-order Clausal Logic)的一个子集。不含递归的DATALOG与关系代数具有相同的表达能力，并且DATALOG语言与关系数据库语言之间有直接的对应关系(见表1)，这为DATALOG作为多关系数据挖掘任务的知识表示方式奠定了良好的基础。

表1 基本概念间的对应关系

关系数据库概念	DATALOG概念
关系名p	谓词符号p
关系p的属性	谓词p的参数
p关系元组$\langle a_1, \cdots, a_n \rangle$	基事实$p(a_1, \cdots, a_n)$
由一个元组集合表示的关系p	由一个基事实集合外延
定义的谓词p定义为视图的关系p	由一个子句内涵定义的谓词p

WARMR使用DATALOG的基事实(Ground Facts)表示多关系数据库中的数据,使用DATALOG查询(DATALOG Queries)表示多关系频繁模式。DATALOG查询具有如下形式:? − A1, A2 ,… , An ,这里的 A_i 是DATALOG原子(DATALOG Atom)。下面我们给出一个例子来说明上述WARMR的知识表示方式。

假设一个多关系数据库由名为custom、parent、buy的三个关系组成,各关系的属性及其元组见表2。

表2 样例多关系数据库

Customer
ID
Allen
Bill
Carol
Diana
Zoe

Parent	
Parent ID	Child ID
Allen	Bill
Allen	Carol
Bill	Zoe
Carol	Diana

Buy	
Custom	IDItem
Allen	Wine
Bill	Cola
Bill	Pizza
Diana	Pizza

原始数据库中的数据由DATALOG的基事实表示,比如关系custom的元组〈Zoe〉相应的基事实为custom(Zoe),parent中的元组〈Allen , Bill 〉相应的基事实为parent(Allen , Bill),buy的元组〈Bill,Pizza〉相应的基事实为buy(Bill,Pizza)。所有数据库中的元组通过这种方式表示为Prolog知识库。

一个DATALOG查询的例子Q1如下:

? -customer(A), parent(A, B), buy(B, pizza)。

DATALOG查询通过一个Prolog引擎提交给Prolog知识库,得到其结果。向一个Prolog知识库r提交一个包含变量{X_1, X_2, …, X_m}的查询Q,? − A1, A2 ,… , An,意味着询问是否存在一个基替代,使得A1, A2 ,… , An形成的合取式在知识库r中为真。查询Q的一个基替代 θ,形为 θ={X_1/c_1, X_2/c_2, …, X_m/c_m},表示将Q中的所有变量 X_i 用相应常数 c_i 置换,得到的置换结果查询用 $Q\theta$ 表示。Prolog引擎对Q的应答结果为使得 $Q\theta$ 在r中成功的所有 θ 形成的集合,这一集合表示为answerset(Q, r)。

为便于理解,这一结果可以用等价地提交给原始多关系数据库的SQL查询来说明。比如,上述例子DATALOG查询Q1等价于如下SQL查询的结果:

```
Select custom. ID, parent. parent ID, parent. child ID,
       buy. custom. ID
From custom, parent, buy
Where custom. ID= parent. parent ID
and parent. child ID = buy. custom. ID
and buy. item= 'pizza'
```

理解了DATALOG查询的语义,我们就可以进一步说明一个DATALOG查询的频度定义了。在传统频繁项集发现中,模式的频度根据事务的标识(Transaction ID)来定义。而在多关系频繁模式发现中,是根据作为发现任务上下文背景中的核心关系的主键值来定义频度。这里的核心关系意味着我们挖掘时关心的核心对象是什么。用ILP技术的基本术语,核心关系描述的是样例(Examples)信息,而非核心关系描述的是与样例有关的背景知识(Background Knowledge)。比如在上述例子数据库中,如果我们的挖掘目标是了解客户的潜在知识,那么custom就是我们的核心关系,custom. ID是custom的主键。parent和buy关系是非核心关系,描述的是custom的相关背景信息。我们可以用如下SQL语句来说明例子查询Q1的绝对频度:

```
Select count(distinct custom. ID)
From custom, parent, buy
Where custom. ID= parent. parent ID
and parent. child ID = buy. custom. ID
and buy. item= 'pizza'
```

上述SQL查询得到的绝对频度与所有客户的数量之比就是查询Q1的相对频度。形式上,在WARMR中通过一个额外的参数(原子key)来实现上述以核心关系为计数方法基础的思想。原子key的谓词符号对应于核心关系的关系名,其谓词参数相应于核心关系的属性。每一个模式中必须包含原子key。这样,一个包含原子key的模式查询Q在知识库r中的相对频度即为

$$freq(Q, r, key) = |\{\theta_k \in answerset(? - key, r) \mid Q\theta_k \ succeeds \ w.r.t. \ r\}| \ / \ |\{\theta_k \in answerset(? - key, r)\}|$$

3.2 WARMR算法

在明确了上述多关系频繁模式发现的基本概念之后,我们分析WARMR算法的核心内容。WARMR算法核心思想与Apriori算法相同,采用自顶向下、逐层、宽度优先的搜索策略。算法迭代执行候选集产生和候选集评估过程。

首先我们讨论WARMR算法的搜索空间。

如果不加限制地使用DATALOG表示模式空间,一方面搜索空间会是无限大;另一方面会导致对于应用领域无意义的模式产生。为了避免这些问题,WARMR算法使用ILP技术中的陈述性语言偏置(Declarative Language Bias)规定合理的模式类型,对搜索空间进行约束。要了解陈述性语言偏置的一般情况,可参见文[19]。具体地,WARMR使用了称之为WRMODE[15]的陈述性语言偏置。WRMODE改编自多关系决策树归纳算法TILDE[20]的语言偏置RMODE。WRMODE规定了可以出现在模式中的原子类型,包括上文所述的key原子,同时规定了原子中的变量的共享性质和具体取值类型。陈述性语言偏置的优点在于用户可以根据对应用领域的理解显式地规定搜索空间,从而WARMR无须改变其具体实现就能够完成不同的复杂任务要求。因此,基于WARMR的系统能够灵活地在一个单一的工具内完成模式类型的改变和相关实验,较之于以前的频繁模式发现算法有更大的灵活性。

为了便于搜索和剪枝,在WRMODE规定的模式空间中,模式按照 θ-包含关系构建搜索空间。θ-包含定义了模式之间的一般与特殊关系。应该注意的是,传统频繁模式发现使用子集关系定义一般与特殊关系,而 θ-包含是一个比子集关系更弱的关系定义。根据 θ-包含整个搜索空间形成一个格(lattice)。

WARMR算法从最一般模式开始,一次搜索 θ-包含格中的一层。算法迭代执行候选集产生和候选集评估过程。表3是WARMR算法的主体部分。步骤(5～10)是算法的主循环。其中步骤(6)使用子算法Warmr-Eval计算每个候选模

式的频度。步骤(7)记录非频繁模式,用于候选模式产生时的搜索空间剪枝。步骤(8)记录频繁模式。步骤(9)使用子算法Warmr-Gen产生候选模式。

表3 WARMR算法

```
Algorithm 1:WARMR
Inputs: Database r; Wrmode language L and key ; threshold minfreq
Outputs: All queries Q ∈L with freq (Q; r; key)≥ minfreq
1. Initialize level d := 1
2. Initialize the set of candidate queries Q1 := {? - key }
3. Initialize the set of infrequent queries I := ∅;
4. Initialize the set of frequent queries F := ∅;
5. While Qd not empty
6.    Find frq (Q; r; key)of all Q∈ Qd using Warmr-Eval
7.    Move the queries∈ Qd with frequency below minfreq to I
8. Update F := F ∪ Qd
9.    Compute new candidates Qd+1 from Qd, F and I using
      Warmr-Gen
10.   Increment d
11. Return F
```

表4是模式频度计算算法Warmr-Eval的描述。值得注意的是,计算过程中使用的是"从解释中学习"(Learning from Interpretation)的策略。在这种策略中,Prolog知识库按照样例分为多个部分,这样在每一层的频度计算时,全部数据只需装入内存一次。与此相应的另一种ILP策略是"从逻辑后承中学习"(Learning from Entailment),该策略对与每个候选模式的评估需要将全部数据装入一次。有关两种策略的试验比较参见文[21]。还应指出的是,文[22]给出了在一定条件下从解释中学习策略的PAC可学习性的证明,这为在关系学习和多关系数据挖掘中广泛、合理地使用从解释中学习策略提供了良好的理论基础。

表4 Warmr-Eval算法

```
Algorithm 2:Warmr-Eval
Inputs: Database r; set of queries Q; Wrmode key
Outputs: The frequencies of queries Q
1. For each query Qj∈Q, initialize frequency counter qj := 0
2. For each substitution θk∈ answerset(? - key, r), do the following:
   (a)Isolate the relevant fraction of the database
      rk⊆ r
   (b) For each query Qj∈Q, do the following:
      If query Qjθk succeeds w. r. t. rk, increment counter qj
3. For each query Qj∈Q, return frequency counter qj
```

表5是产生候选模式子算法Warmr-Gen。根据θ-包含关系的反单调性,非频繁模式的特化模式也是非频繁的。因此在候选集产生阶段,首先根据陈述性语言偏置WRMODE,仅产生当前层频繁模式的最一般特化,实现搜索空间的初步剪枝。进一步地,频繁模式的特化也可能是已知非频繁模式的特化或者可能与已知频繁模式等价,因此这类特化也应剪枝,其剪枝操作分别对应于步骤(2. i)与步骤(2. ii)。

表5 Warmr-Gen算法

```
Algorithm 3 : Warmr-Gen
Inputs: Wrmode language L; infrequent queries I; frequent queries F; frequent queries Qd for level d
Outputs: Candidate queries Qd+1 for level d+1
1. Initialize Qd+1 :=∅;
2. For each query Qj∈ Qd, and for each immediate specialization
Qj' ∈ L of Qj:
Add Q'j to Qd+1, unless:
(i)Qj' is more specific than some query ∈ I, or
(ii)Qj' is equivalent to some query ∈ Qd+1∪F
3. Return Qd+1
```

3.3 多关系频繁模式发现ILP方法的进展

在分析了算法的基本概念和核心内容之后,我们有必要指出的是,WARMR虽然奠定了多关系频繁模式ILP方法的基础,但是从实践的层面看,WARMR算法在效率和可扩展性方面的性能并不理想。这主要是因为,WARMR在候选模式评估和产生阶段的计算极大地依赖θ-包含关系,而θ-包含计算实质上是一个NP完全问题[23]。在多关系频繁模式ILP方法的后续发展中,如何在WARMR方法基础上提高计算效率和可扩展性成为研究重点之一。

一般地,一个多关系频繁模式发现系统的时间消耗主要由三个部分组成:候选模式产生阶段时间消耗、候选模式评估阶段的时间消耗,以及数据从外存向内存重复装入的时间消耗。哪一个因素是系统的性能瓶颈,取决于所使用系统的方法特征、应用领域的特征以及输入参数的具体情况。但根据我们的分析和试验,一般情况下,候选模式评估阶段的时间消耗经常占据系统整体时间消耗的主要部分,85%以上的比率并不少见,因此,提高候选模式评估阶段的性能,成为改善多关系频繁模式发现系统效率和可扩展性的重点。候选模式产生阶段时间消耗主要来源于有关候选模式的基于θ-包含的等价与特化关系测试计算,因此通过设计一定的方法,消除上述测试的必要性是提高系统效率的方法之一。对于数据装入带来的性能问题,如3.2节所述,从解释中学习的ILP策略对于一般数据集合能够很好地解决问题,但对于样例规模普遍大于主存的多关系数据库,这一方法的有效性还有待进一步的研究和发展。

2000年,文[24]通过使用称之为"查询包"(Query Packs)的技术来提高候选模式评估阶段的性能。试验结果表明,在一定情况下,较之于经典WARMR算法的原始方法,该方法性能可以提高一个甚至多个数量级。使用查询包策略改善了标准的Prolog执行策略,不是一次评估一个候选模式,而是能够一次实现一个候选模式集合的评估,从而提高了评估效率。查询包有效利用了搜索空间形成的格的特征,将具有一定类似结构的候选模式组织成一个查询实体,从而新颖地实现了共享计算,减少了相似候选模式间的冗余计算。应该注意的是,这一方法实现于底层Prolog引擎级,不是上层算法级。

2001年,文[25]提出了FARMER算法。FARMER算法是比WARMR算法效率更高的一个多关系频繁模式发现算法。FARMER的效率提高主要在于在候选模式阶段避免了模式间的等价测试。FARMER通过引入一个特殊的数据结构trie,结合冗余受限的陈述性语言偏置,产生候选模式,保证了算法不会产生相互之间等价的模式。数据结构trie同时也用于候选模式的频度计数,从而系统不再依赖于Prolog引擎。FARMER算法的一个缺陷在于其语言偏置有特殊的规定,因而一定程度上影响了模式表达的能力和灵活性。

2003年,FARMER算法的提出者在文[26]中进一步对FARMER算法提出了修正,使得θ-包含关系下的等价测试问题转换为"目标一致"(Object Identity)关系下的等价测试问题。"目标一致"关系的引入,一方面消除了原始FARMER算法中语言偏置的有关限制,另一方面构建了候选模式的有效搜索路径,为等价测试和搜索空间剪枝提供了优化基础。试验表明,在小规模数据集上改进后的FARMER算法模式,表达力相当于WARMR算法,效率高于WARMR算法一个甚至多个数量级。

但我们在这里必须指出的是,改进前后的FARMER算法都需要使用树结构把全部频繁模式以及部分非频繁模式存储于主存中。我们的试验表明,对于内在蕴涵复杂大规模搜

索空间的应用,该算法可能会在搜索过程中耗尽主存空间而终止。因此,FARMER 算法效率虽然比 WARMR 算法提高了,但其可扩展性仍有待进一步研究。

2004 年,文[27]提出了 RADAR 算法,使用文档检索系统中的倒排索引(Inverted Index)技术来优化模式评估阶段的性能。倒排索引是一种基于磁盘的搜索结构。RADAR 算法使用这一数据结构,在数据预处理阶段首先将原始多关系数据库扁平化。在挖掘阶段,候选模式产生与评估采用深度优先的方式,候选模式评估不再使用 Prolog 引擎,而是基于压缩后的倒排索引结构,从而提高了效率和可扩展性。对于大规模多关系数据库,较之于 WARMR 算法与 FARMER 算法,RADAR 算法在可扩展性方面具有优势。

多关系频繁模式发现 ILP 方法的研究进展,除了效率和可扩展性方面的研究,还涉及到频繁模式从一般形态向更特殊形态的演化。下面我们介绍分析这方面的主要成果。

2001 年,文[28] 首先提出了多层多关系频繁模式发现方法。2004 年相同的作者在文[29]完善了这一方法。多层多关系频繁模式发现方法可以看作是传统多层频繁模式发现方法[30]向多关系数据挖掘领域的更新。在文[29]中,研究者将描述逻辑(Description Logics)与 ILP 技术结合,完成涉及多个概念层的频繁模式发现任务。方法使用 AL-log 作为知识表示语言。AL-log 是一种混合型语言,集成了描述逻辑 ALC[31]和 DATALOG 语言。描述逻辑 ALC 能够按照概念、规则与个体描述结构化知识。个体表示应用领域的对象,概念表示对象的类别,规则表示概念间的二元关系。复杂概念可以使用基础概念和规则建构得到。通过描述逻辑 ALC 形成的知识库由内涵和外延两部分组成,内涵部分通过 is-a 关系建立概念层次,规定概念间的包含关系,而外延部分说明个体与概念间的实例关系。在 AL-log 中,ALC 的作用实质上是根据应用领域结构化知识对 DATALOG 部分的变量和常数进行限定和说明,因而 AL-log 能够同时描述结构化与关系性知识。另外,在搜索空间的构造方面,方法基于"一般化包含"(Generalized Subsumption)关系[32]而不是 θ-包含。一般化包含能够利用应用领域的先验知识确定模式之间的一般与特殊关系。在 AL-log 知识库中包含应用领域的结构化先验知识,一般化包含关系能够根据这些先验知识,精化产生对应用有趣的模式,有效地限制搜索空间。多层多关系频繁模式发现方法实现于空间数据多层关联规则挖掘 ILP 系统 SPADA 中,取得了良好应用。应该指出的是,基于一般化包含关系的计算较之于 θ-包含更具复杂性,因此对于文[29]中提出的方法,在计算效率和可扩展性方面的研究还有较大的空间。

2004 年,文[33]拓展了一部分传统频繁模式中的精简化概念,定义了新颖的、面向语义和数据的、一阶逻辑下的闭合频繁模式与自由频繁模式概念。基于这些概念的频繁模式发现方法,可以根据应用领域的先验背景知识挖掘没有冗余的模式集合。在算法的效率方面,该方法也借鉴了上述效率和可扩展性方面的历史研究成果并且使用了一个面向频繁模式发现的优化 Prolog 引擎。试验结果表明,在效率和发现结果的有趣性方面,该方法较之于原始 WARMR 方法有较大的提高。对于大规模数据集,我们认为可扩展性方面的性能问题尚需更多的试验和理论分析。

4 多关系频繁模式发现非 ILP 方法

多关系频繁模式发现研究领域存在着并非基于 ILP 技术的方法。从知识表示方式的角度,这些方法可以分类为基于"选择图"的多关系数据挖掘方法[34]和基于图的数据挖掘方法[35~45]。我们认为,从知识表示方式的角度来看,这些方法的模式表示语言表达能力一般不如基于 ILP 技术的方法,对于复杂的多关系挖掘领域基于 ILP 技术的方法应该是重点发展方向;但是从算法效率、应用领域扩展和方法论角度来看,这些方法提供了新颖有效的技术和思想,为 ILP 方法提供了有益的补充,扩展了多关系频繁模式发现方法谱系,更为重要的是为系统深入地探究多关系频繁模式发现的本质特性和提出优化发现方法提供了更广泛的归纳分析内容。因此,我们也介绍与分析这些方法的重点研究成果。

本文 4.1 节分析基于"选择图"的方法;4.2 节分析基于图的数据挖掘方法。

4.1 基于"选择图"的方法

1999 年,文[34]提出了一个多关系数据挖掘框架和计算体系结构。

在其框架中,模式被看作是对多关系数据库中一组对象的内涵式表达,这种表达可以用 SQL 语句表示。为了有效地建构搜索空间,模式进一步由与 SQL 语句等价的选择图(Selection Graphs)表示和构建。一个选择图是有向图,图的一个节点表示多关系数据库中的一个关系和施加于该关系属性上的选择条件集合,边表示所连接的两个节点对应的关系之间的连接条件,一般为两关系之间的主外键关系。基于选择图的模式精化可以通过增加节点内的属性条件约束、增加节点、增加边,或同时增加节点和边实现。在加边或同时增加节点和边过程中,可以利用关系数据库的数据库模式(Database Scheme)所提供的主外键信息或其他关系连接信息,有效压缩搜索空间,避免无趣模式的产生,提高算法的效率。

选择图可以较容易地转换为 SQL 形式。利用该框架提供的各种基于 SQL 的原语,可以方便有效地得到模式评估过程中的各种统计参数。

在计算体系结构方面,使用了客户机/服务器的形式。客户机完成空间搜索和候选模式的产生,而服务器完成候选模式的评估。客户机/服务器体系结构一定程度上实现了分布计算,提高了算法性能。

整体上,该方法与 ILP 方法的不同在于:一方面是模式表示语言的不同,另一方面是模式评估方式的不同。这两方面的区别使得该方法无需如同 ILP 方法那样需要 Prolog 引擎:在数据预处理阶段无需将原始数据库表示为 Prolog 知识库,并且可以直接向驻留在服务器上的关系数据库管理系统(RDBMS)提交 SQL 语句,即可完成模式评估。系统能够充分地利用关系数据库管理系统的数据组织存储优化能力、查询优化能力,从而提高了数据挖掘算法的效率和扩展性。

文[34]仅提出了基于上述框架和体系结构的多关系频繁模式发现算法思想,并没有提出算法的具体实现和试验情况,因而这一方法在多关系频繁模式发现领域的进一步研究还有较大的空间。但是,其提出的与 RDBMS 耦合的思想及其有效利用数据库模式(Database Scheme)信息的方法,对于多关系频繁模式发现其他方法是具有借鉴意义的。

4.2 基于图的数据挖掘方法

基于图的数据挖掘方法从结构化数据拓扑结构的角度研究数据挖掘。数学上最一般的拓扑结构是图。XML 与 HTML 文本、符号序列、树以及多关系数据等复杂结构化数据都可以用图结构表达。图结构数据在大量的现实领域广泛存

在,比如生物学、化学、材料科学以及通信网络等。从图结构表达的数据中发现具有共性的子结构是数据挖掘的重要任务和方法。频繁子结构的发现最终将从结构化数据中识别出某些概念,这些概念能够描述有趣的结构模式。被发现的子结构概念是对细节数据结构的抽象并提供解释数据的相关属性。基于图的数据挖掘方法以频繁子结构发现为任务,研究相关方法及其应用问题。因为图结构数据在实际应用领域中通常存储于多关系数据库,所以基于图的数据挖掘方法可以看作为多关系数据挖掘的一个领域。

基于图的数据挖掘方法将结构化数据库表达为图,在图上执行数据挖掘任务。图作为离散数学研究最一般的数据结构之一,包括了从不同角度来看具有不同特征的子结构。在基于图的数据挖掘中,子结构被分为不同的类别,包括一般子图、归纳子图、连接子图、树(有序或无序)、路径。

这方面的研究最早的成果出现于1994年,文[35]提出的SUBDUE能够发现有效压缩原始图的特征子图。在此之后,提出了许多基于图的数据挖掘方法与系统。1994年,文[36]提出的GBI能够发现图数据中的频繁子图。他们的方法都使用贪婪搜索机制,以避免图同构计算带来的计算复杂性,因此最终得到的子图模式集合是不完全的。

在2000年,文[37,38]提出了AGM算法,该算法融合了Apriori算法和数学图论的思想,发现图数据中的频繁子图和关联规则。AGM的基本原理类似于Aprioi算法,一个图由一个事务组成,最终产生关联规则,采用完全的自底向上的层级搜索。AGM能够处理有向无向、无标和有标图,能够挖掘的子图模式包括一般子图、归纳子图、连接图、有序子树、无序子树、子路径;图数据被转化为一个邻接矩阵,不需要太大的计算量;为了提高效率,有标图的节点标识和边标识用自然数索引;在实际实现上,邻接矩阵用"码"表示,根据码的情况,方法引入了有序码和有序表的概念。

以AGM为基础,许多研究继续以层级搜索和数学图论思想的融合作为基本原则开展工作。2001年,文[39]提出的FSG方法在算法效率上优化了AGM。2002年,文[40]提出的gFSG,能够从数据集中枚举所有几何子图。2002年,文[41]提出的gSpan,使用DFS编码构建有序标识,在内存消耗和计算效率上优于前期工作提出的方法。2003年,文[42]提出的CloseGraph,不是直接挖掘完全集,而是挖掘闭合子图。

有别于AGM上述研究线索的研究重点描述如下:

2001年,文[43]提出了基于版本空间(Version Space)的算法MolFea,该算法从图数据中发现特征路径。2003年,文[44]提出的FFSM,使用一个代数图框架来解决子图同构问题。2004年,文[45]提出的Gaston,能有效地挖掘频繁路径。

这里需要指出的是,基于ILP的方法也能够挖掘具有图结构的数据[16]。本质差异在于,ILP方法使用的知识表示方式是DATALOG,而基于图的数据挖掘方法使用的知识表示方式是数学图。另外,ILP方法发现的模式更具复杂性,而基于图的数据挖掘方法效率更高。如何结合两者各自的优势,设计能够高效发现复杂模式的多关系频繁模式方法是将来研究发展的重点之一。

总结 随着各种数据中知识发现的深入发展,数据挖掘的对象和发现的模式的复杂程度不断增长。对于复杂多关系数据,传统的数据挖掘方法难以完成相应的任务。多关系频繁模式发现研究致力于发现复杂关系数据库中涉及多个关系的复杂频繁模式,从上世纪90年代以来取得了蓬勃发展,并成功地应用于各种等领域。相对于国际发展态势,国内学界在多关系频繁模式发现领域的介绍和研究并不多见,这有别于传统数据挖掘的研究状况。本文期望将来这种状况有所改变。

最后,我们指出多关系频繁模式发现研究面临的挑战和需要重点解决的问题:

1)传统频繁模式发现特殊方法和特殊任务向多关系领域的全面扩展和更新。

2)能够在一个统一框架下完成单关系和多关系频繁模式发现的方法及其理论基础。

3)对于大规模海量数据集的可扩展性方法研究。

4)基于适宜代数理论的多关系频繁模式发现算法与关系数据库系统和数据仓库系统的无缝连接方法。

5)时变、多源、多类型数据的多关系频繁模式发现方法。

6)融合ILP方法与基于图的数据挖掘方法的各自优势,设计能够高效发现复杂模式的多关系频繁模式方法。

7)多关系频繁模式发现方法向关系-对象数据和面向对象数据的拓展。

参考文献

1 Agrawal R, Srikant R. Fast algorithms for mining association rules in large databases. In: Proc. of 20th Intl Conf. on Very Large Databases, Santiago, Chile, 1994

2 Hipp J, Guntzer U, Nakaeizadeh G. Algorithms for Association Rule Mining - A General Survey and Comparison. In: Proc. ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining, 2000

3 Dzeroski S, Lavrac N. Relational Data Mining. Springer, Berlin, 2001

4 Dzeroski S, De Raedt L, Wrobel S. Proc. of the First Intl Workshop on Multi-Relational Data Mining. Eighth ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining, Edmonton, Canada, 2002

5 Dzeroski S, De Raedt L, Wrobel S. Proc. of the Second Intl Workshop on Multi-Relational Data Mining. Ninth ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining, Washington, DC

6 Dzeroski S, De Raedt L, Wrobel S. Proc. of the Third Intl Workshop on Multi-Relational Data Mining. Tenth ACM SIGKDD Intl Conf. on Knowledge Discovery and Data Mining, Seattle, WA

7 De Raedt L. Attribute-value learning versus Inductive Logic Programming: the Missing Links (Extended Abstract). In: Proc. of the 8th Intl Conf. on Inductive Logic Programming. vol 1446 of Lecture Notes in Artificial Intelligence, Springer-Verlag, 1998

8 Blockeel H. Top-down induction of first order logical decision trees. Artificial Intelligence, 1998, 101: 285～297

9 Getoor L. Multi-relational data mining using probabilistic relational models: research summary. In: Proc. of the First Workshop in Multi-relational Data Mining, 2001

10 Wrobel S. Inductive Logic Programming for Knowledge Discovery in Databases. In :[1]. pages 74～101

11 Lavra_c N, D_zeroski S. Inductive Logic Programming: Techniques and Applications. Ellis Horwood, Chichester, 1994

12 De Raedt L. Logical settings for concept-learning. Artificial Intelligence, 1997,95(1)

13 Muggleton S, De Raedt L. Inductive logic programming : Theory and methods. Journal of Logic Programming, 1994,19,20:629～679

14 Dehaspe L, De Raedt L. Mining association rules in multiple relations. In: Proc. of the 7th Intl Workshop on Inductive Logic Programming. vol 1297 of Lecture Notes in Artificial Intelligence, 1997

15 Dehaspe L, Toivonen H. Discovery of Frequent Datalog Patterns. Data Mining and Knowledge Discovery, 1999, 3(1):7～36

16 Dehaspe L, Toivonen H, Kind RD. Finding frequent substructures

in chemical compounds. In:Proc. of the KDD-98, 1998
17 Mannila H, Toivonen H. Levelwise search and borders of theories in knowledge discovery. Data Mining and Knowledge Discovery, 1997, 1(3):241～258
18 Ullman J D. Principles of Database and Knowledge-Base Systems. volume I. Rockville, MD: Computer Science Press, 1988
19 Nedellec C, Ade H, Bergadano F, et al. Declarative bias in ILP. In:De Raedt L, ed. Advances in Inductive Logic Programming, volume 32 of Frontiers in Artificial Intelligence and Applications, IOS Press, 1996. 82～103
20 Blockeel H, Raedt L D. Top-Down Induction of First-order Logical Decision Trees. Artificial Intelligence, 1998, 101(1-2): 285～297
21 Blockeel H, De Raedt L, Jacobs N. Scaling up ILP by learning from interpretations. Data Mining and Knowledge Discovery, 1999, 3(1):59～93
22 De Raedt L, Dzeroski S. First order jk-clausal theories are PAC-learnable. Artificial Intelligence, 1994, 70:375～392
23 Kietz J U, Lubbe M. An Efficient Subsumption Algorithm for Inductive Logic Programming. In: Proc. of the Eleventh Intl Conf. on Machine Learning, Morgan Kaufmann Publishers, Inc, San Mateo, CA, 1994. 130～138
24 Blockeel H, Dehaspe L, Demoen B, et al. Executing Query Packs in ILP. In: Proc. of the l0th Intl Conf. on Inductive Logic Programming, volume 1866 of Lecture Notes in Artificial Intelligence, Springer, London, UK, 2000. 60～77
25 Nijssen S, Kok J N. Faster Association Rules for Multiple Relations. In: Proc. of the 17th Intl Joint Conf. on Artificial Intelligence, Morgan Kaufmann Publishers, Inc, Seattle, USA, 2001. 891～896
26 Nijssen S, Kok J N. Efficient frequent query discovery in FARMER. In :13th Intl. Conf. on Inductive Logic Programming (ILP 2003), 2003
27 Clare A, Williams H E, Lester N M. Scalable Multi-Relational Association Mining. In:Proceedings of the 4th International Conference on Data Mining ICDM '04, 2004
28 Malerba D, Lisi F A. An ILP method for spatial association rule mining. In:Working notes of the First Workshop on Multi-Relational Data Mining, Freiburg, Germany, 2001
29 Lisi F A, Malerba D. Inducing Multi-Level Association Rules from Multiple Relations. Mach Learn, 2004, 55, 2 : 175～210
30 Han Jiawei, Fu Yongjian. Mining Multiple-level Association Rules in Large Databases. IEEE Transactions on Knowledge and Data Engineering, 1999, 11(5)798～805
31 Schmidt-SchaubB M, Smolka G. Attributive concept descriptions with complements. Artificial Intelligence, 1991, 48(1):1～26
32 Buntine W. Generalized subsumption and its applications to induction and redundancy. Artificial Intelligence, 1988, 36(2):149～176
33 De Raedt L, Ramon J. Condensed Representations for Inductive Logic Programming. In: Proc. of Ninth Intl Conf. on the Principles of Knowledge Representation and Reasoning, 2004
34 Knobbe J, Blockeel H, Siebes A, et al. Multi-relational Data Mining. In:Proc. of Benelearn '99, 1999
35 Cook J, Holder H. Substructure discovery using minimum description length and background knowledge. J Artificial Intel Research, 1994, 1:231～255
36 Yoshida K, Motoda H, Indurkhya N. Graph-based induction as a unified learning framework. J of Applied Intel, 1994, 4:297～328
37 Inokuchi A, Washio T, Motoda H. An Apriori-based Algorithm for Mining Frequent Substructures from Graph Data. In: Proc. of the 4th European Conf on Principles of Data Mining and Knowledge Discovery, 2000. 13～23
38 Inokuchi A, Washio T, Motoda H. Complete Mining of Frequent Patterns from Graphs: Mining Graph Data. Machine Learning, 2003, 50: 321～354
39 Kuramochi M, Karypis G. Frequent Subgraph Discovery. In: Proc. of the 1st Intl Conf. on Data Mining, 2001
40 Kuramochi M, Karypis G. Discovering frequent geometric subgraphs. In:ICDM, 2002. 258～265
41 Yan X, Han J. gSpan: Graph-based Substructure Pattern Mining. In: Proc of the 2nd Intl Conf on Data Mining, 2002
42 Yan X, Han J. Closegraph: mining closed frequent graph patterns. In:KDD, 2003
43 De Raedt L, Kramer S. The levelwise version space algorithm and its application to molecular fragment finding. In:IJCAI'01: Seventeenth Intl Joint Conf on Artificial Intelligence. vol2, 2001. 853～859
44 Huan J, Wang W, Prins J. Efficient mining of frequent subgraphs in the presence of isomorphism. In:ICDM, 2003. 549～552
45 Nijssen S, Kok J N. A quickstart in frequent structuremining can make a difference. In:KDD, 2004. 647～652

(上接第 157 页)

参 考 文 献

1 王建会，申展，胡运发. 一种实用高效的聚类算法. 软件学报, 2004, 15(5): 697～705
2 Hearst M A, Pedersen J. Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results. In: Proc. of the 19th Annual Int'lACM/SIGIR Conf. Zurich, 1996. 76～84
3 Willet P. Recent Trends in Hierarchic Document Clustering: A Critical Review. Information Processing and Mangement, 1988, 24(5): 577～597
4 Rocchio J J. Document Retrieval Systems—Optimization and Evaluation :[PhD dissertation]. Harvard University, Cambridge, MA, 1966
5 Cutting D R, Pedersen J O, Karger D R, et al. Scatter/Gather: A Cluster-based Approach to Browsing Large Document Collections In: Proc. of the 15th Annual Int'l ACM/SIGIR Conf. Copenhagen, 1992. 318～329
6 Xu Jian Suo. Wang Li. TCBLHT: A New Method of Hierarchical Text Clustering. In: Proceedings of 4th International Conference on Machine Learning and Cybernetics, 2005. 2178～2181
7 Dumais ST, Furnas GW, Landauer TK, et al. Using Latent Semantic Analysis to Improve Information Retrieval. In: Proceedings of CHI'88, 1988. 281～285
8 Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases. In: Proceedings of the 20th International Conference on Very Large Data Bases, 1994. 487～499
9 Antonie M, Zaiane O R. Text Document Categorization by Term Association. In:Proc. of IEEE Intl. Conf. on Data Mining, 2002, 19～26
10 Meretakis D, Fragoudis D, Lu Hongjun, et al. Scalable Association-based Text Classification. In:Proc. of the 2000 ACM CIKM International Conference on Information and Knowledge Management 2000, 6～11
11 Aberdeen J, Burger J, Day D, et al. MITRE: description of the Alembic system used for MUC-6. In:Proceedings of the 6th Message Understanding Conference, 1993. 141～155
12 Porter M F. An algorithm for suffix stripping. Program, 1980, 14(3):130～137
13 Roberto J, Bayardo Jr. Efficiently Mining Long Patterns from Databases. In: Proceedings ACM SIGMOD International Conference on Management of Data, 1998. 85～93
14 Salton G, Buckley B. Term-Weighting approaches in automatic text retrieval. Information Processing and Management, 1988, 24(5): 513～523
15 Ziarko W. Variable precision rough set model. Journal of Computer and System Sciences, 1993, 46(1): 39～59