结构关系模式挖掘算法测试平台的设计与实现①

An Algorithm-Test Platform for Structural Relation Pattern (SRP) Mining

陈未如　吴　迪　张　雪　(沈阳化工学院 计算机科学与技术学院 辽宁 沈阳 110142)

摘　要： 建立了一个结构关系模式挖掘算法测试平台。平台应用组件技术，把测试算法封装到组件中，并为测试算法提供一个通用的接口，使得算法能以组件的形式嵌入测试平台。平台提供数据处理、算法试验、结果的可视化与评价等模块，使用户评价新算法的过程变得更加客观、简便。实验结果表明，借助该测试平台，既可以方便的对各种算法进行测试，又有助于方便的研究新的算法。

关键词： 结构关系模式挖掘 算法测试平台 组件技术 数据挖掘

1　引言

数据挖掘(Data Mining)，又称为数据库中的知识发现(Knowledge Discovery in DataBase,KDD)，就是从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[1]。

许多从事数据挖掘算法研究的人员在学习、研究中都会有这样的体会，为了验证一个新算法的性能，从数据准备到算法结果评价都要自己设计完成，整个过程繁琐，有相当大的难度[2]。

随着数据挖掘技术的发展，国内外出现了形形色色的数据挖掘工具，其中一部分提供了二次开发接口，如：(Weka，ARMiner)，用户可以在此基础上开发自己的算法，并集成到系统中去运行，达到测试算法的目的。与前面提到的方法相比，这种方法减少了工作量，但它也存在一些缺陷，这些软件在设计上都是倾向于挖掘任务，而不是算法测试，要完成用户各种各样的算法测试相对比较困难。

为解决算法测试中存在的这些问题，建立了一个专门用于算法性能测试的算法测试平台，此平台集成了数据处理、算法试验、结果的可视化与评价等功能，使得算法开发人员不再将大量精力集中在一些重复性工作上，而把精力都投入到算法本身的研究和实现上。

2　结构关系模式挖掘算法测试平台的特点

序列模式挖掘是一种非常重要的数据挖掘技术，结构关系模式挖掘是在序列模式挖掘基础上提出的一种新的数据挖掘任务，又叫做后序列模式挖掘。结构关系模式是一种包括有序关系、并发关系、互斥关系、重复关系及这些关系的复合关系的模式[3]。目前，结构关系模式挖掘研究已取得初步成果，提出了一些定义、性质及定理，同时也提出了一些相关的挖掘算法。为了能够便捷、准确的对这些结构关系模式挖掘算法的性能进行评价，本文研究一种结构关系模式挖掘算法测试平台。

结构关系模式挖掘算法测试平台应用组件技术给算法提供了一个通用的接口和基于 COM 标准的算法公共代码框架。用户只需将算法封装成 COM 组件，就可以很容易地动态将算法嵌入到实验平台中进行测试。平台提供数据处理、算法测试、测试结果的可视化及评价等模块，同时提供算法库和结果库来保存算法测试信息，方便日后用户进行查询和比较。

3　设计原则和功能要求

结构关系模式挖掘算法测试平台是一个能帮助算法研究人员快速、简便测试新算法性能的工具。平台体系结构见图 1。

① 基金项目:辽宁省教育厅科学研究计划(05L338)
收稿时间:2008-11-06

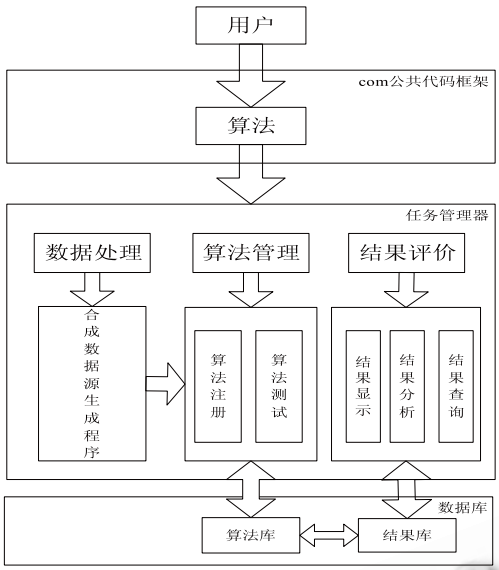

图 1 算法测试平台体系结构

平台提供了基于 COM 标准的算法公共代码框架，方便用户对编写好的挖掘算法进行封装。任务管理器是系统的核心，用户编写的算法在此进行注册、测试得到测试结果，并对测试结果进行分析。同时平台提供算法库和结果库来保存算法测试的相关信息。

4 主要功能模块及特点

4.1 数据处理

对算法进行测试必须要有测试数据。测试所需数据源分为合成数据源和真实数据源，使用真实数据的问题是数据本身不够规则，常常含有噪声，而数据挖掘工作仅仅是整个 KDD 工作中的一个阶段[4]，前面还有许多预处理和数据转换工作，挖掘算法本身已经不考虑(或者很少考虑)数据中的噪声等问题，因此本测试平台可以提供合成数据源[5]，通过其所具有的数据可控性来检测算法的正确性与效率。利用这种可控性将希望算法能够发现的知识“埋藏”在生成的合成数据中，通过观察算法在找寻这些知识时的表现，判断算法是否正确以及效率如何。

4.2 算法管理

(1) 算法注册：在算法注册模块进行算法注册，即填写一张算法注册表，用以记录算法的各种属性信息，包括算法的名称、类型、存储位置、运行环境、参数说明等信息，完成算法库的建立。

(2) 算法测试：组件化的测试方法，为了方便用户进行算法测试，平台采用组件技术来实现对算法的封装，使得算法能以组件的形式进入测试平台，通过平台提供的接口来传递参数信息并获取运行结果。

4.3 结果评价

此模块的功能是对算法运行结果进行显示，并给与适当的分析。算法只有经过结果评价才可以确认其实用性，只有运行结果优秀的算法才有存在的意义。

(1) 结果显示：采用文本形式显示算法测试结果。

(2) 结果分析：结果的分析离不开对分析结果的可视化，可视化技术使用户可以对算法的性能有直观的认识[6,7]。平台提供了文本和图形两种表示方式。结果分析方法有两个方面，一是单一算法的性能分析，二是同类算法的性能比较。具体的分析形式有如下三种情况：

① 相同算法，相同数据集，不同参数,算法运行时间比较。

相同算法，相同数据集，不同参数，算法产生总模式个数比较。

② 相同算法，相同参数，不同数据集,算法运行时间比较。

相同算法，相同参数，不同数据集,算法产生总模式个数比较。

③ 不同算法，相同数据集，相同参数,算法运行时间比较。

不同算法，相同数据集，相同参数,算法产生总模式个数比较。

(3) 结果查询：对算法测试结果进行保存，通过平台提供的结果查询功能，用户可以查询注册算法的所有运行结果。

5 COM公共代码框架

为了便于用户编写自己算法的 COM 组件，平台提供公共代码框架，即算法接口规范文档，用户只要按此将算法功能实现和封装，使得通过接口平台可以调用并运行算法即可。算法组件模型，如图 2 所示。算法对应着三个重要接口，其中 IAlgInput 为算法的输入接口，用来接收测试算法的输入信息；第二个接口 IAlgOutput 为算法的输出接口，用来返回算法的测试结果；第三个接口 IAlgRes 为算法资源接口，用来计算算法所需要的资源。在 COM 规范中，采用 IDL 来定义接口内容，下面是具体的接口定义[8]：

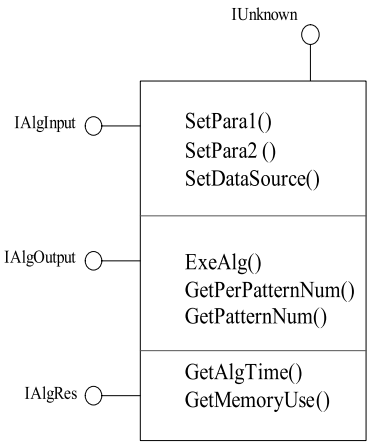

图 2　算法组件接口

(1) 算法输入接口 IAlgInput

Interface IAlgInput:Iunknown

{

HRESULT SetPara1([in]double Para,[out]char [256] message);

//当算法要求有一个参数输入时，选择此方法。

HRESULT SetPara2([in]double Para1,[in] doublePara2,[out]char[256]message);

//当算法要求有两个参数输入时，选择此方法。

HRESULT SetDataSource([in]BSTR SourcePath，[out]char[256] message);

//输入数据源文件的路径，数据源格式为文本类型。

};

参数说明：

Para，Para1，Para2：表示算法需要的参数值。

SourcePath ：表示数据源文件的路径。

Message：若成功则无意义(保留)，若失败返回错误信息。

(2) 算法输出接口 IAlgOutput

Interface IAlgOutput:Iunknown

{

HRESULT ExeAlg([out]BSTR ResultName);

//运行算法。

HRESULT GetPatternNum([in]BSTR ResultName,[out]int *PNum);

//获得结果文件中的总模式个数。

};

参数说明：

ResultName ：结果文件路径及名称。结果文件格式为　文本类型。

PNum ：结果文件中的总模式个数。

(3) 算法资源接口 IAlgRes

Interface IAlgRes:Iunknown

{

HRESULT GetAlgTime ([out]ULONG AlgTime);

//输出算法处理时间。

HRESULT GetMemoryUse([out]ULONGMemoryUse);

//输出内存占用量

};

参数说明：

AlgTime：返回算法运行时间。

MemoryUse：返回内存占用量。

6　实验与分析

利用平台对重复关系模式挖掘算法进行实验分析，重复关系模式挖掘现有算法包括：基于序列模式挖掘的重复序列模式挖掘，基于最大序列模式集的重复序列模式挖掘，基于最大序列模式集的最大重复序列模式挖掘以及基于客户序列数据库的重复序列模式挖掘等。

以基于序列模式的重复序列模式挖掘算法为例，数据集 C10-T8-S8-N1k-D10k 做为输入数据源，其中数据集中涉及到的参数具体说明如下表 1：

表 1　数据集参数

符号	描述
|D|	事务数据的数量
|C|	每个客户平均事务数
|T|	事务数据平均长度
|S|	序列数据平均长度
|N|	不同项目的个数

数据集由平台集成的 IBMQest 数据源生成程序生成，所有参数由用户通过界面设定，在此数据集下运行算法，通过设置不同参数得到算法测试结果如表 2.

表 2 实验测试结果

最小重复度 (%)	重复模式个数 (个)	挖掘时间 (秒)
5	21	0.05
4.5	24	0.12
4	30	0.27
3.5	38	0.36
3	55	0.67
2.5	97	0.92
2	183	1.23
1.5	423	2.85
1	1228	5.04
0.5	10997	9.36

对测试结果进行分析，采用单一算法分析方法，以数据集做为分析参数，分析最小重复度与生成重复模式个数间的关系，结果分析的图形化表示如图 3。由分析结果可以清晰的得出结论重复模式个数随着最小重复度的增加而递减。

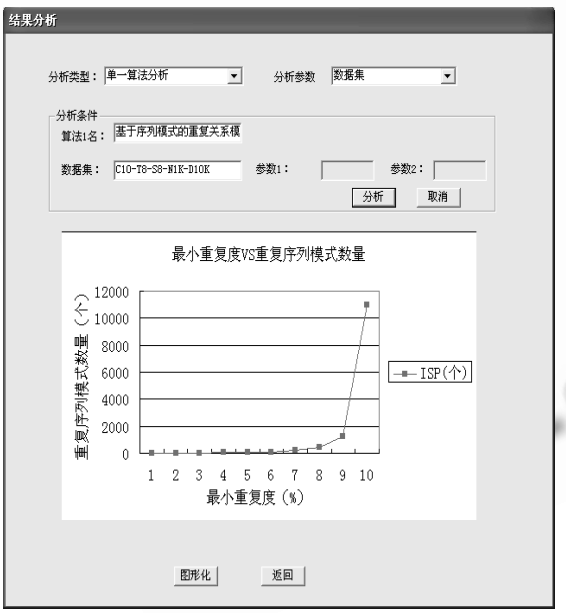

图 3 实验结果分析

7 结论

主要介绍了结构关系模式算法测试平台的体系结构和基于 COM 标准的算法公共代码框架。测试平台以组件为桥梁，为测试算法的性能提供了一个有效平台。算法研究人员可以使用这个平台对其所研究的算法进行性能测试，可以将其算法与其它同类算法进行比对。平台为算法的实现提供了一个统一的框架，对于已经实现了的算法程序，还可以通过编写适配器的形式嵌入到平台中。通过实验初步验证了该平台的优越性和可行性。

参考文献

1 邵峰晶,于忠清.数据挖掘原理与算法.北京:中国水利水电出版社:1－30.

2 Feelders A.Briefings Methodological and practical aspects of data mining.Elsevier Science, 2000,37:271－281.

3 Lu J, Adjei O, Wang XF, Hussain F. Sequential Patterns Modeling and Graph Pattern Mining. Proceedings of the Tenth International Conference IPMU,Perugia,Italy, 2004,2.

4 Piatetsky-Shapiro G, Fayyad U, Smith P. From data mining to knowledge: an overview. Advances in Knowledge Discovery and Data Mining.Cambridge Mass:AAA/MIT Press, 1996:1－34.

5 纪元,陈未如,张雪.并发关系模式合成数据源生成方法.山东大学学报(理学版), 2007,42(9):84－87.

6 Ling XC, Huang J, Zhang H. AUC: A Statistically Consistent and More Discriminating Measure than Accuracy. Proc. IJCAI’03, 2003:329－341.

7 Huang J, Ling CX.Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering, 2005,17:299-310.

8 Rogerson D. COM 技术内幕—微软组件对象模型.北京:清华大学出版社, 1999:20－38.