基于质量的数据挖掘服务选择

李玉华　陈云开　卢正鼎

（华中科技大学计算机科学与技术学院　武汉 430074）

摘　要　在面向服务的数据挖掘系统中各种数据挖掘的算法封装成 Web 服务。用户选择合适的数据挖掘服务执行自己的数据挖掘任务，而大多数最终用户并不具备这样的专业知识。从方便用户的角度出发，系统需提供一套服务选择机制，来帮助用户选择高质量的数据挖掘服务。综合通用 Web 服务的评价标准、数据挖掘领域的专用评价因子及用户评价反馈等多种因素及服务的动态性，给出了一个较全面的数据挖掘服务评价本体，讨论了服务质量的评价方法，给出了基于服务质量评价的动态数据挖掘服务选择方法，用户可根据数据挖掘服务评价本体的语义模型，输入质量约束条件，也可以调整评价因子权值，系统在满足用户约束条件的服务集中，通过计算出服务的综合质量值，挑选最适合的算法执行。

关键词　服务质量，数据挖掘，服务选择

The User-centered Data Mining Ontology Development on Universal Knowledge Grid

LI Yu-Hua　CHEN Yun-Kai　LU Zheng-Ding

(College of Computer Science, Huazhong University of Sci. & Tech., Wuhan 430074)

Abstract　In the service-sriented data mining system, the Data Mining (DM) algorithms are packed to Web services. User can define DM task by selecting proper Data Mining Service (DMS), but most users haven't such professional knowledge. One service selection mechanism is needed to help user selecting high quality DMS in view of user usability. A more all-around DMS Quality Evaluation ontology(OntDMQ) is proposed by synthesizing Web service quality of services (QoS), DM unique characteristic, subjective factor such as user feedback and service dynamic characteristic. The evaluation method of QoS is discussed. The QoS-based dynamic service selection method is presented that user can define the QoS constraint of DMS referencing OntQE and adjust the factor fff, the system select the most appropriate DMS in the services fiting the user requirements according to computing compositive quality value.

Keywords　Quality of services, Data mining, Service selection

近年来，关于在面向服务的体系结构（Service Oriented Architecture，SOA）上提供知识服务渐渐成为了研究的热点。在面向服务的数据挖掘系统中采用 Web Service 的标准技术来描述和发现所有可能的数据挖掘算法。在整个系统中，各种数据挖掘的算法封装成 Web 服务。随着服务数量和种类的增加，用户选择空间增大，针对同一种服务需求可供选择的服务会越来越多。

数据挖掘服务是涉及数据、计算、挖掘知识的复杂服务应用，用户需要具备非常全面的专业知识才能正确使用和选择。而大多数最终用户并不具备这样的专业知识，从方便用户的角度出发，需提供一套服务选择机制，来帮助用户选择高质量的数据挖掘服务。

综合通用 Web 服务的评价标准、数据挖掘领域的专用评价因子及用户评价反馈等多种因素及服务的动态性，在不同的时间服务的质量不同，给出了一个较全面的数据挖掘服务评价本体。

用户可根据数据挖掘服务评价本体的语义模型，输入质量约束条件，也可以调整质量因子权值（重要程度）。系统在满足用户约束条件的服务集中，通过计算出服务的综合质量值，选择最适合的算法执行。

1　数据挖掘服务评价本体

数据挖掘是一种 Web 服务，因此数据挖掘服务评价本体（OntDMQ）要包括公共 Web 服务质量评价因子要求和数据挖掘领域的专用评价因子。Web 服务质量有很多文献进行讨论和研究[1~6]，综合前人研究的成果，结合数据挖掘领域的特点，提出了较全面的综合了主、客观因素数据挖掘服务评价本体，如图 1 所示。若将本体中的专用评价因子，替换成其他领域的评价因子，则形成其他领域服务评价本体，具有较好的通用性。数据挖掘服务评价本体主要包括公共评价因子（common factor），专用评价（special factor）因子和过程评价因子（process factor）。

数据挖掘服务评价本体的每个概念类包括 9 个主要属性描述：ClassName，weight，Haschild，Value Type，Effect Type，EvaluateMethod，EstimateFunction，NodeValue，Unit。

ClassName 为该概念类名称，以概念名称为惟一标志，各服务评价因子之间不允许有重名。

weight 是评价因子的权值，同时评价本体上节点的权重有以下约束条件：

（1）根节点的权重为 1；

（2）任意一个评价因子节点 s 的权重是它的所有子节点权重的总和。

Haschild 标明是否有子节点，若有子节点，则其本身没有独立的指标值，由其子指标共同表征。

Value Type 代表取值类型（数值型、区间型、语言型、等级型）。

Effect Type 代表因子的质量影响，其中效益型因子如可用性，标明指标值越大越好，而成本型因子如价格，指标值越

小越好。

EvaluateMethod 标明指标取值方法(固定型、统计型、计算型、设定型)；EstimateFunction 为计算性指标的估算函数。NodeValue 为评价因子的取值。Unit 为指标取值的单位。

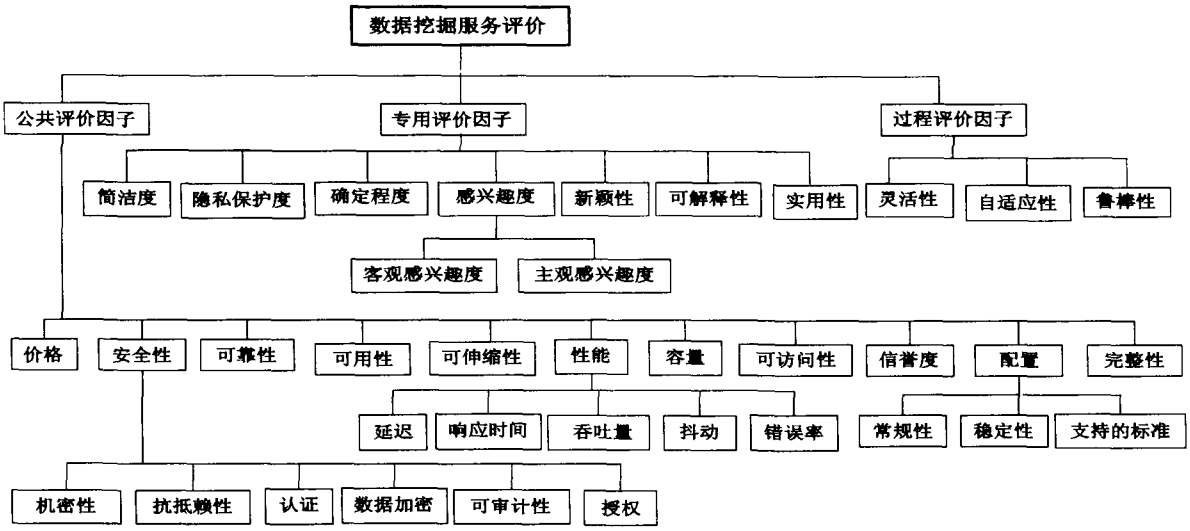

图 1　数据挖掘服务评价本体

1.1　公共评价因子(Common factor)

包括价格、可用性、可访问性、可伸缩性、完整性、可靠性、性能、容量、配置安全性、信誉度。分别定义如下：

价格(Cost)：价格的计算方法为服务提供者给出。

可用性(Availability)：可用性是指服务是否存在或是否已就绪可供立即使用。可用性表示服务可用的可能性。可用性的取值范围[0,1]，较大的值表示服务一直可供使用，而较小的值表示无法预知在某个特定时刻服务是否可用。与可用性有关的还有修复时间(time-to-repair，TTR)。TTR 表示修复已经失效的服务要花费的时间。理想情况下，较小的 TTR 值是合乎需要的。

可访问性(Accessibility)：可访问性表示能够为服务请求提供服务的程度。它可以表示为一种可能性尺度，用来表示在某个时间点上成功地实例化服务的成功率或机会。可访问性不同于可用性，服务可用但却无法访问这种情形是可能存在的。可以通过构建一个可高度伸缩的系统使服务得到很高的可访问性。

可伸缩性(Scalability)：可伸缩性描述在一定时间段内可增加计算能力以处理更多用户请求、操作或事务的能力。可伸缩性是指不管请求量如何变化，都能够始终如一地为请求服务的能力，即算法对于大规模海量数据的良好的处理能力[7]。

数据挖掘有两种可伸缩性问题：行(数据库大小)可伸缩性和列(维)可伸缩性。如果一个数据挖掘系统行数扩大了 10 倍，而执行同样的数据挖掘查询的时间最多也不超过其原来时间 10 倍的话，则说这个系统是行可伸缩的；如果数据挖掘查询执行时间和列(属性或维)数呈线性增长关系，则说这个系统是列可伸缩的。由于多维性的原因，使一个数据挖掘服务成为列可伸缩的比让其成为行可伸缩的更具有挑战性。

完整性(Integrity)：完整性指服务如何维护交互相对于最初情况的正确性。适当地执行服务事务会实现正确的交互。一个事务是指一系列将被当作单个工作单元的活动。要使事务成功，必须完成所有的活动。如果一个事务未完成，那么所做的全部更改都被回滚。Web 服务内容、数据资源、SOAP 信息不会被未授权的用户篡改或破坏。数据的完整性是一个布尔型的参数，即服务是否支持数据完整性。

可靠性(Reliability)：可靠性表示能够维护服务和服务质量的程度。在另一种意义上，可靠性是指服务请求者和服务提供者发送和接收的消息的有保证和有序的传送。度量可靠性的参数包括平均修复时间 MTTR 和失效间平均时间 MTBF。

性能(Performance)：性能包括吞吐量(Throughput)、延迟(Latency)、响应时间(ResponseTime)、抖动(Jitter)和错误率(ErrorRate)。

延迟是发送请求和接收响应之间的往返时间。主要从通信、计算和知识集成的角度来衡量，通信时间主要依赖于操作模型和体系结构及网络性能如带宽和延迟有很大关系。在 SOA 结构中，通信时间为将算法软件移动到远程数据集所在机器的时间。计算时间是一个独立于模型的参数，主要由算法选择来决定。知识集成时间是将各地的结果集成起来所花费的时间。

通常高性能的服务应该是提供高吞吐量、低延迟、快速响应时间、低抖动和低错误率。

容量(Capacity)：容量是指服务能同时处理的最大用户请求数。

配置(Configuration)：服务配置与接口更新程序或采用的标准有关。它提供服务兼容的标准信息，它指示服务间是否组合。包括稳定性(Stability)、支持的标准(SupportedStandard)和常规性(Regulatory)。稳定性表示服务接口变化频度。支持的标准描述服务兼容的标准。Web 服务使用许多标准，例如 SOAP、UDDI 和 WSDL。常规性指服务与规则、法律一致，遵循标准和已建立的服务级别协议的可能性。要正确调用服务请求者请求的服务，就必须严格遵守服务提供者所提供的正确版本的标准。

安全性(Security)：安全性是服务质量的一个方面，通过验证涉及到的各方、对消息加密以及提供访问控制来提供机密性(Confidentiality)、可审计性(Auditability)、认证(Authentication)、授权(Authorization)、数据加密(Data Encryp-

tion)和不可抵赖性(NonRepudiation)。由于服务调用是发生在公共的因特网上,安全性的重要性已经增加。根据服务请求者的不同,服务提供者可以用不同的方法来提供安全性,所提供的安全性也可以有不同的级别。

信誉度(reputation):信誉度是服务可信度的一个度量,由用户投票选出。它主要取决于终端用户使用服务的经验。对于相同的服务不同的用户有不同的意见。信誉度的值定义为用户对服务评价的平均值。

$$Q_{REP}=\sum_{1}^{n}R_i/n \tag{1}$$

式中 R_i 是用户的服务评价,n 是服务被评价的次数,通常用户在某个取值范围内对服务评价。如在 Amazon. com,评价范围是[0,5]。

1.2 专用评价因子(Special factor)

对于数据挖掘应用域而言,专用的评价因子主要是评价数据挖掘算法所挖掘的模式的评价以及对算法性能的评价,由于数据挖掘可挖掘数据特征化、概念描述、关联分析、分类、预测、聚类分析、链结分析、孤立点分析、演变分析等多种模式,对于每类可挖掘的模式,评估模式兴趣度有效性的方法也有所不同。数据数据挖掘系统具有产生数以千计、甚至数以万计的模式或规则的潜在能力。"所有模式都是有趣的吗?"答案是否定的。实际上,对于给定的用户,在可能产生的模式中,只有一小部分是他感兴趣的。一个模式是有趣的(interesting),如果(1)它易于被人理解;(2)在某种程度上,对于新的或测试数据是有效的;(3)是潜在有用的;(4)是新颖的。如果一个模式符合用户确信的某种假设,它也是有趣的[1]。有趣的模式表示知识。

对于数据挖掘算法的评价,有很多文献进行了研究和讨论。文[7]对于形如 $X\Rightarrow Y$ 的关联规则,可用支持度,置信度进行客观兴趣度的度量。通过简洁性、确定性、实用性和新颖性的模式兴趣度评估数据挖掘的模式。数据挖掘对聚类的典型要求包括可伸缩性、处理不同类型属性的能力、发现任意形状的聚类、用于决定输入参数的领域知识最小化、处理噪声数据的能力、对于输入记录的顺序不敏感、对于输入记录的顺序不敏感、高维性、基于约束的聚类、可解释性和可用性等。文[8]指出预测的准确率、速度、鲁棒性、可伸缩性、可解释性是评估分类和预测方法的五条标准。

文[9]提出了一个 KDS 的质量本体 KDSQO,从通用测度、专用测度、过程测度三个方面对 KDS 进行评价,但没有考虑服务的安全性,在专用测度中也没有考虑隐私保护度,在专用测度中,只考虑规则的测度,不够全面。

对于数据挖掘服务而言,一个公有的专用评价因子就是隐私保护度[10,11]。尽管客观度量可以帮助识别有趣的模式,但是仅有这些还不够,还要结合反映特定用户需要和兴趣的主观度量。

综合以上的研究成果,确定专用评价因子包括简洁性、确定性、实用性、新颖性、感兴趣度、可解释性、可视化、隐私保护度。模型强健性等可归为过程评价因子。

1. 简洁性(Simplicity)

模式兴趣度的一个重要因素是对于人的理解,模式的总体简洁性。模式简洁性的客观度量可以看作模式结构的函数,用模式的二进位位数,或属性数,或模式中出现的操作符数来定义。

规则长度是一种简洁性的度量,规则的长度越小,那么其紧致性就越好。对于用合取范式(即合取谓词的集合)表达的规则。规则的长度简单地定义为规则中合取符的个数。关联、判别或分类规则的长度超过用户定义的阈值时,被认为是不感兴趣的。对于以判定树表达的模式,简洁性可以是树叶或树节点的个数的函数。

形如 $X\Rightarrow Y$ 规则的简洁度可以用如下公式计算:

$$simplicity(X\Rightarrow Y)=1-\frac{Headsize}{Maxsize} \tag{2}$$

其中,$HeadSize$ 是前提 X 中条件的数量,$Maxsize$ 是前提 X 中的条件的最大数。

2. 确定性(Certainty)

每个发现的模式都应当有一个表示其有效性或"值得信赖性"的确定性度量。

对于形如 $X\Rightarrow Y$ 的关联规则,其确定性度量是置信度(confidence),置信度是条件概率 $P(Y/X)$ 即包含 X 的事务也包含 Y 的概率。更形式地,置信度定义为:

$$confidence(X\Rightarrow Y)=P(Y/X) \tag{3}$$

研究表明[12],如下的计算方法相对于 confidence 而言有着更优的效果,计算方法如下:

$$certa\ int\ y(X\rightarrow Y)=\max\left[\frac{P(X/Y)-P(Y)}{1-P(Y)},\frac{P(X/Y)-P(X)}{1-P(X)}\right] \tag{4}$$

其中,P 代表概率,max 是求最大值函数。

对于分类规则,式(4)也可以方便地作为可靠性或准确性的确定性的度量。对于聚类而言,就可以用聚类的质量来度量,聚类的质量是基于对象相异度来评估的,相异度可以对多种类型的数据来计算,包括区间标度变量、二元变量、标称变量、序数型变量和比例标度型变量,或者这些变量类型的组合。

3. 实用性(Utility)

一个模式的潜在的有用性是定义其兴趣度的一个重要因素。它可以用一个实用性函数(如支持度)来评估。关联模式的支持度是模式为真的任务相关的元组(或事务)所占的百分比。对于形如 $X\Rightarrow Y$ 的关联规则,支持度是概率 $P(X\cup Y)$,其中 $X\cup Y$ 表示同时包含 X 和 Y 事务,即项集 X 和 Y 的并。支持度定义为 $\sup port(X\Rightarrow Y)=P(X\cup Y)$。

特征和判别描述基本上是广义元组。

4. 新颖性(Novelty)

新颖的模式是那些提供新信息或提高给定模式集性能的模式。例如,一个数据异常可以认为是新颖的,它不同于根据统计模型和用户的信念所期望的模式。检测新颖性的另一策略是删除冗余模式。如果发现的规则被已在知识库中或导出的规则集中的另一规则所蕴涵,则两个规则都要重新审查,以便去掉潜在的冗余。

5. 感兴趣度(Interest)

实际上,感兴趣度应该从主观和客观两个方面来衡量。

(1)客观感兴趣度。经典的客观感兴趣度是在文[13]中描述的 Piatesky-Shapiro 测度,在文[14]中给出了一个性能更优的变种。

$$int\ erest(X\Rightarrow Y)=\sqrt{PS(X\Rightarrow Y)\times\frac{P(X/Y)}{N}} \tag{5}$$

其中,PS 代表 Piatesky-Shapiro 测度为 $P(X,Y)-P(X)P(Y)$,P 代表概率,N 是数据集 F 的大小,即总纪录数。

(2)主观感兴趣度。尽管客观度量可以帮助识别有趣的

模式，但是仅有这些还不够，还要结合反映特定用户需要和兴趣的主观度量。主观兴趣度度量基于用户对数据的确信。这种度量发现有趣的模式，如果它们是出乎意料的(对照用户的确信)，或者提供用户可以采取行动的策略信息。

6. 可解释性(Interpretability)

数据挖掘模式的可理解性。对于分类而言，涉及学习模型提供的理解和洞察的层次。对于聚类而言，用户希望聚类结果是可解释的，可理解的和可用的。也就是说，聚类可能需要和特定的语义解释和应用相联系。应用目标如何影响聚类方法的选择也是一个重要的研究课题。

7. 可视化(Visuality)

“一幅图胜过一句话”，在数据挖掘中是非常真实的。数据挖掘的可视化分为数据可视化，挖掘结果可视化，挖掘过程可视化和可视化数据挖掘。数据挖掘服务可视化质量和灵活性严重地影响了数据挖掘服务的使用、解释能力和吸引力。

8. 隐私保护度(Privacy protection)

尽管数据挖掘应用对社会目标有很大利用价值，但对它的使用引发了保留意见。数据挖掘对隐私造成威胁，数据挖掘可以将不同的数据集关联起来进行私人信息的分析，另一方面根据利用数据挖掘工具所得出的推断来解释、应用和采取相应的行动。被挖掘的一般模式将人们分为不同的种类，并以一定的概率暴露了个人的机密3信息。如何在保证隐私的情况下挖掘出有用的信息是近年来数据挖掘领域研究的热点之一，很多学者研究了基于隐私保护的数据挖掘算法。

隐私保护度是数据挖掘服务质量评价的重要部分，文[15]利用转移概率矩阵对原始数据进行变换，实现数据隐私保护。给出了隐私保护程度的量化表示。将属性的隐私保护程度表示为“数据变换后，该属性不同取值个数”除以“原始数据中，该属性不同取值个数”，计算公式为“该属性转移概率矩阵中不等于零的元素个数”的平方根除以“该属性原始数据不同取值个数”。

1.3 过程评价因子(Process factor)

从过程控制的角度出发，服务还应该是灵活的、自适应的和鲁棒的。

1)灵活性(flexibility)。灵活性指的是应该可以灵活处理各种不同情况。如对于聚类分析，处理不同类型属性的能力，发现任意形状的聚类，对于输入记录的顺序不敏感，发现基于约束的聚类等。

2)鲁棒性(Robustness)。而面对数据中的各种不完全性(如数据缺失或语义不完整)，服务的鲁棒性是质量评价所需考虑的特性。

3)自适应度(Adaptiveness)。自适应度指的是服务能够自动判断所需的信息，并利用系统环境自动寻找，并将辅助信息融入到数据挖掘服务中。所谓“自适应”就是能够根据不同的情况，自动快速调节，对原有行为或其他一些东西进行改变，以满足新情况新环境的要求！

2 本体实现

数据挖掘服务评价本体是一个树状结构，包括数据挖掘评价、公共评价因子、专用评价因子、过程评价因子、价格、安全性、可靠性、可用性、可访问性、常规性、信誉度 、互操作性、性能等概念类。数据挖掘评价是Web服务的综合服务质量，和公共评价因子、专用评价因子、过程评价因子概念间是子类关系(即subclassof关系)，公共评价因子、专用评价因子、过程评价因子与其下级的价格、可用性、可访问性、可伸缩性、完整性、可靠性、性能、容量、配置安全性、信誉度和简洁性、确定性、实用性、新颖性、感兴趣度、可解释性、可视化、隐私保护度等节点间也是子类关系。安全性又包括机密性、数据加密等子类，权值、节点值等是机密性概念的属性。

在本体的编码中采用W3C提出的OWL[16](Web Ontology Language)语言。本体开发工具采用Stanford的protege2000。

图2中代码是基于OWL的本体描述的一部分。

```
<owl:Ontology rdf:about=""/>
 <owl:Class rdf:about="#QosFactor">
   <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"
   >这是基类</rdfs:comment>
 </owl:Class>
<owl:Class rdf:about="#CommonFactor">
   <rdfs:subClassOf>
     <owl:Class rdf:ID="QosFactor"/>
   </rdfs:subClassOf>
 </owl:Class>
...
<owl:Class rdf:ID="Scalability">
   <rdfs:subClassOf rdf:resource="#CommonFactor"/>
 </owl:Class>
...
 <owl:Class rdf:ID="Security">
   <rdfs:subClassOf>
     <owl:Class rdf:about="#CommonFactor"/>
   </rdfs:subClassOf>
 </owl:Class>
   <owl:Class rdf:ID="Confidentiality">
   <rdfs:subClassOf rdf:resource="#Security"/>
 </owl:Class>
...
 <owl:ObjectProperty rdf:ID="Weight">
   <rdfs:domain rdf:resource="#Confidentiality"/>
 </owl:ObjectProperty>
...
 <owl:ObjectProperty rdf:ID="NodeValue">
   <rdfs:domain rdf:resource="#Confidentiality"/>
 </owl:ObjectProperty>
 </rdf:RDF>
```

图2 数据挖掘服务评价本体的OWL的代码片断

3 基于质量的数据挖掘服务选择

3.1 数据挖掘服务的描述及发现

系统需要服务提供者提供两个独立的WSDL文档：实现文档和接口文档。实现文档主要描述服务的具体实现机制。在接口文档中，每一个DMS的描述分为两个部分：功能和质量部分。功能部分描述了该服务在执行时所需的元数据，而质量部分描述了该服务在不同数据集上的质量表现。图3是可视化链接分析服务的评价因子，只是给出了评价参数的类型、取值和单位。

图4是可视化链接分析服务的WSDL形式描述片断。

Accessibility 0.99000	Security
Availability 0.99995	Audiability 1
Performance	Authentication “Password”
(max) Jitter 1 (msec)	Authorization “SSL”
(max) ErrorRate 10^{-5}	Confidentiality 1
(max) Latency 300 (msec)	DataEncryption “AES-128”
(min) Throughput 384 (Kbps)	NonRepudiation 1
(max) ResponseTime 0.01 (sec)	Simplicity 0.6
Cost 300 (RMB/minute)	Certainty 0.7
Capacity 200	Utility 0.85
Scalabillity 0.80	Novelty 0.8
Configuration	Interpretability 0.82
SupportedStandards “UDDI 3.0”	Interest 0.87
SupportedStandards “WSDL 1.1”	Visuality 1
Reliability	Privacy protection 0
MTBF 36,000,000 (sec)	
MTTR 1,800 (sec)	

图3 可视化链接分析服务质量评价因子

```
〈?xml version = "1. 0" encoding = "U TF28" ?〉
〈definitions name = "Apriori"
xmlns = "http :/ / schemas. xmlsoap. org/ wsdl/ "
xmlns : soap = "http :/ / schemas. xmlsoap. org/ wsdl/
soap/ "
xmlns :tns = "http :/ / localhost/ Apriori"
xmlns :xsd = "http :/ / www. w3. org/ 1999/ XMLSchema" 〉
〈port Type name = "SLAGM_port" 〉
〈operation name = "SLAGM"
…
〈/ operation〉
〈/ port Type〉
〈Quality Description〉
〈Data Characteristic〉
〈Number of records〉 13594 〈/ Number of records〉
…
〈Ratio of partially specified data〉 3. 07 % 〈/ Ratio
of partially
specified data〉
…
〈/ Data Characteristic〉
〈Certainty〉 0.92 〈/ Certainty〉
〈Response time〉 33.7 Sec 〈/ Computation time〉
…
```

图 4　数据挖掘服务的 WSDL 描述片断

具体的控制流程简要描述如下：客户端应用基于 HTTP 协议来向服务器端传送 SOAP 格式的请求。以 Java Servlets 形式运行的 UDDI 作为一个服务器端进程服务来接受客户端应用请求，并以 SOAP 格式返回结果。

3.2　服务质量评价

1. 收集服务质量信息

QoS 指标取值并非静态不变的，服务选择需要参考动态预测的指标值。按照指标取值随时间变化及对服务上下文的依赖程度，选择不同的服务质量信息收集方式。可以将指标取值类型分为：①固定型，由服务提供方在服务描述中设定或更新；②统计型，在每次服务执行结束时由提供方或中介服务重新计算更新；③计算型，与服务运行上下文（客户状态、客户需求或提供方状态）之间有固定的函数依赖关系；④设定取值：对于确定取值的评价因子，如价格由服务提供者在服务注册时设定，设定型，其对服务执行上下文的依赖过于复杂而难以建模，需要人工或利用软件工具进行预测设定。

在实际的执行监控中收集信息：在每次服务执行结束时由提供方或中介服务重新计算更新。实际的服务的执行历时由服务请求者收集。这需要所有服务请求者使用接口来实现一些机制来测量服务历时。

从用户反馈中收集质量信息：每个用户都要求更新他/她刚使用的服务的质量信息，这样服务质量的注册对所有的用户都公平，因为 Web 服务的质量值是基于真实用户的经验。为了避免服务质量被某一个团体操纵，用户在更新反馈前必须获得一个服务提供者授权的一对密钥。

2. 评价数据的标准化

因为 OntDME 的评价因子有不同类型的取值范围，所以为了保证评价因子的权重具有可比性，需要通过评价数据的预处理操作，将所有评价因子的值域调整到一个统一的区间。因为多数评价因子的评价效果与它的值域具有线性变化的特点，所以系统主要采用最小-最大规格化方法来规格化评价数据。

假设在一次评价过程的多组数据实例中，$\max(v_i)$ 是评价因子 f_i 取值的最大值，$\min(v_i)$ 是评价因子 f_i 取值的最小值。对于效益型因子则根据式(6)进行规格化，例如 Web 服务的信誉度；对于成本型因子则根据式(7)进行规格化，比如 Web 服务的响应时间。经过数据规格化后，任何一个评价因子 f_i 的取值范围由原来的$[\min(v_i), \max(v_i)]$区间转换到$[0,1]$区间。

$$v_i' = \begin{cases} \dfrac{\max(v_i) - v_i}{\max(v_i) - \min(v_i)}, & \max(v_i) - \min(v_i) \neq 0 \\ 1, & \max(v_i) - \min(v_i) = 0 \end{cases} \tag{6}$$

$$v_i' = \begin{cases} \dfrac{v_i - \min(v_i)}{\max(v_i) - \min(v_i)}, & \max(v_i) - \min(v_i) \neq 0 \\ 1, & \max(v_i) - \min(v_i) = 0 \end{cases} \tag{7}$$

对于枚举类型的评价因子，则需要领域专家同时给出每种枚举类型的质量评分映射表，系统根据这张映射表来完成评价因子从枚举型到数值型的类型转换过程。

3. 质量评价

经过预处理后的数据实例，每个 DataItem 项的值域都在$[0,1]$区间内，因此可以限定每个评价因子的权重也在$[0,1]$区间内。

设定评价本体有 n 个评价因子$\{f_1, f_2, \cdots, f_n\}$，每个评价因子 $f_i (i=1 \sim n)$的权重为 w_i，服务实例 SI 的一个数据实例 DI 在每个评价因子上的取值为$\{v_1, v_2, \cdots, v_n\}$，那么服务实例 SI 的质量是 DI 在每个评价因子上取值的加权总和，下面给出评价的计算公式：

$$Quality(SI) = \sum_{i=0}^{n} v_i w_i \tag{8}$$

由式(8)可以看出，数据挖掘服务质量的评价结果决定于评价因子的取值和评价因子的权重分布。评价因子的取值来源于 Web 服务的运行环境，各影响因子的权值的设定有三种方式：服务发布时可由服务提供者根据经验设定，用户选择服务时可对服务评价因子的权值进行调整，当有很多用户使用后，可根据用户使用记录的信息利用机器学习的方法进行修正。

4. 动态服务选择的过程

动态服务选择的过程如下：

(1)用户可输入质量因子约束，也可对系统推荐的质量评价因子权值进行调整。

(2)系统根据用户输入的质量约束，先删除不满足约束的待选服务。

(3)对剩余的服务则根据服务综合指标的计算方法，计算出待选服务的综合质量值并进行排序。

(4)系统选择满足质量约束条件的最优的服务，完成用户的数据挖掘任务，同时将结果返回给用户。

4　服务选择的实例

采用 Java 语言实现了反洗钱应用中的若干算法，包括可视化链接分析，基于可疑帐户的链接分析，基于约束的非指导性链接发现方法，频繁子图发现，基于图熵的链接发现等以及 Dijkstra 算法，PFS 算法、two-tree PFS[17] 用于交易网络分析的算法。对于每个算法都封装成 Web 服务，进行服务描述包括其质量信息，在注册中心上进行注册，服务质量信息一部分根据算法实际性能设定，另外一些根据经验设置，如价格、可靠程度、可用程度等参数。

用户提出的应用需求问题，某一个可疑账户的"交易网络分析"，质量需求价格小于 388 元。系统解析后提供 8 个待选

数据挖掘服务，待选服务指标信息如表1待选服务指标信息所示。

表1 待选服务指标信息

服务	价格(cost)	可视化(visuality)	响应时间(秒)	…
可视化链接分析	310	1	0.12	…
基于可疑帐户的链接分析	360	0	0.1	…
基于约束的非指导性链接发现方法	400	0	0.5	…
频繁子图发现	320	0	0.4	…
基于图熵的链接发现	280	0	0.8	…
Dijkstra算法	370	0	0.2	…
PFS算法	340	0	0.3	…
two-tree PFS	320	0	0.16	…

在算法选择时，首先过滤那些价格大于或等于388的服务，根据服务评价本体计算综合评价，如表2所示待选数据挖掘服务质量综合评价结果，最终系统选择综合评价最优的可视化链接分析服务并执行。

表2 待选数据挖掘服务质量综合评价结果

服务	综合评价值
可视化链接分析	0.681
基于可疑帐户的链接分析	0.567
频繁子图发现	0.559
基于图熵的链接发现	0.512
Dijkstra算法	0.206
PFS算法	0.369
two-tree PFS	0.537

结束语 在面向服务的数据挖掘系统中，从方便用户的角度出发，提供一套服务选择机制，来帮助用户选择高质量的数据挖掘服务是十分必要的。文章给出了基于数据挖掘服务质量的服务选择方法。首先讨论了数据挖掘服务评价本体的设计，综合Web服务质量研究成果和数据挖掘服务的特点，提出了较全面的综合了主、客观因素的数据挖掘服务评价本体，数据挖掘服务评价本体主要包括公共评价因子，专用评价因子和过程评价因子，给出各个评价因子的定义。讨论了数据挖掘服务评价本体的实现。然后讨论了基于质量的数据挖掘服务选择方法。最后讨论了反洗钱领域数据挖掘服务选择的实例。

参考文献

1 Yang Lei, Dai Yu, Zhang Bin, et al. A dynamic Web service composite platform based on QoS of services. in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Harbin, China: Springer Verlag, Heidelberg, D-69121, Germany, 2006. 709～716

2 Tsesmetzis D T, Roussaki I G, Papaioannou I V, et al. QoS awareness support in Web-Service semantics. In: Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, AICT/ICIW'06. Guadelope, French Southern Territories: Institute of Electrical and Electronics Engineers Computer Society, Piscataway, NJ 08855-1331, United States, 2006. 128～134

3 Zhou Chen, Chia Liang-Tien, Lee Bu-Sung. DAML-QoS ontology for Web services. In: Proceedings - IEEE International Conference on Web Services, ICWS 2004. San Diego, CA, United States: IEEE Computer Society, Los Alamitos; Massey University, Palmerston, CA 90720-1314, United States; New Zealand, 2004. 472～479

4 Conti M, Kumar M, Das S K, et al. Quality of Service Issues in Internet Web Services. IEEE Transactions on Computers, 2002, 51(6): 593～594

5 Liu Bixin, Wu Quanyuan, Jia Yan, et al. QoS aware service composition with multiple quality constraints. in: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Beijing, China: Springer Verlag, Heidelberg, D-69121, Germany, 2005. 123～131

6 Luo Junzhou, Ji Peng, Wang Xiaozhi, et al. A novel method of QoS based resource management and trust based task scheduling. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Xiamen, China: Springer Verlag, Heidelberg, D-69121, Germany, 2005. 21～32

7 Han Jiawei, Micheline K. Data Mining Concepts and Techniques. Morgan Kanfmann Publishers, 2001. 3～4

8 M.马斯蒂安，武森，高学东. 数据仓库与数据挖掘. 超星数字图书馆，2003. 234～235

9 杨立，左春，王裕国. 面向服务的知识发现体系结构研究与实现. 计算机学报，2005，28(4)：445～457

10 Bertino E, Fovino I N, Provenza L P. A framework for evaluating privacy preserving data mining algorithms. Data Mining and Knowledge Discovery, 2005, 11(2): 121～154

11 Zhang Nan, Zhao Wei, Chen Jianer. Performance measurements for privacy preserving data mining. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Hanoi, Viet Nam: Springer Verlag, Heidelberg, D-69121, Germany, 2005. 43～49

12 Delgado M, Snchez D. Martn2Bautista M J, et al. Mining association rules with improved semantics in medical databases. Artificial Intelligence in Medicine, 2001, 21(1): 241～245

13 Silverstein A, Brin S, Motwani R. Beyond market basket s: Generalizing association rules to dependence rules. Data Mining and Knowledge Discovery, 1998, 2(1): 39～68

14 Tan P, Kumar V, Srivastava J. Selecting the right interestingness measures for association patterns. In: Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining, Canada, Edmonton, Alberta, Canada, 2002. 32～41

15 葛伟平，汪卫，周皓峰，等. 基于隐私保护的分类挖掘. 计算机研究与发展，2006，43(1)：39～45

16 OWL Web Ontology Language Overview. http://www.w3.org/2004/OWL/ (2004)

17 Xu J, Chen H. Fighting Organized Crime: Using Shortest-Path Algorithms to Identify Associations in Criminal Networks. Decision Support Systems, 2004, 38(3): 473～487