联系发现:一种新的数据挖掘方法综述*)

陈 飞 商 琳 骆 斌 陈世福

(南京大学计算机软件新技术国家重点实验室 南京 210093)

摘 要 联系发现是数据挖掘中较新的研究领域。联系发现是一种对海量数据进行挖掘,找出其中潜在模式,抽取有用知识并发现隐藏联系的技术。本文首先综述了联系发现的概念、范围、特点和难点等,详细介绍了联系发现的几种主要方法:无监督的联系发现方法(新颖联系发现)、使用归纳逻辑程序技术挖掘关联数据的联系发现方法、多假设反演推理的联系发现方法、基于相关分析的联系发现方法以及KOJAK组队探测器,讨论了联系发现系统性能评估的方法与联系发现的置信区间度量方法,并简要描述了联系发现的一个具体应用的实例——证据抽取和联系发现研究计划(EELD),最后探讨了目前联系发现研究中出现的问题及未来发展趋势。

关键词 联系发现,数据挖掘,反演推理,相关分析,性能评估

Link Discovery: A New Data Mining Method

CHEN Fei SHANG Lin LUO Bin CHEN Shi-Fu

(National Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093)

Abstract Link Discovery is a relatively new form of data mining. Link Discovery is the technique for mining large amounts of data to find hidden patterns, extract valuable knowledge and discover hidden links. A survey on link discovery is presented in the paper, including the definition, the research range, characteristics and challenges of Link Discovery. And several typical Link Discovery methods are detailed, including unsupervised Link Discovery methods (novel Link Discovery), relational data mining with inductive logic programming for Link Discovery, multi-hypothesis abductive reasoning for Link Discovery, Link Discovery based on Correlation Analysis and the KOJAK group finder. We emphasize on the performance evaluation metrics for Link Discovery systems and measuring confidence intervals in Link Discovery. Further more the aplication of Link Discovery is presented: EELD. Lastly some issues about Link Discovery are discussed and the future directions of Link Discovery are pointed out.

Keywords Bayesian learning, Reinforcement learning, Single-agent, Multi-agent

1 引言

联系发现,也可称为连接发现(link discovery),是数据挖掘中较新的研究领域,被认为是关系挖掘的挑战之一[1]。联系发现是对海量数据进行挖掘,找出其中潜在的模式,抽取其中有用的知识,发现其中隐藏的联系的技术。联系发现的目的是从海量的异构数据集合中,自动识别出异常和危险的活动。成功的联系发现应用有:发现隐藏的组织结构或关联群体、识别欺骗行为、模仿团队行为以及尽早探测出新的威胁[2]。Mooney等人认为:联系发现是从大量的关系数据中识别出可能的潜在威胁活动的模式,且这些模式是已知的、复杂的以及多关系的[3]。Sentor把联系发现描述为:发现已知模式的证据,更主要的是发现未知的但可能是很重要的联系[4]。联系发现的范围很广,包括社会关系分析、欺骗调查、图理论、模式分析和联系分析(Link Analysis)等[5]。Han和Kamber认为,联系发现的挖掘方法不同于传统的数据挖掘方法[6]。联系发现的一般过程是:首先产生关联模式,随后从多个数据库中查出匹配给定模式的实例[7]。联系发现研究综合了多个领域的知识,包括:离散数学(图论)、社会学(社会网络分析)和计算机科学等。目前,联系发现已广泛应用于许多社会领域,如:法律诉讼调查、反欺诈侦察、网络分析以及电信通讯等[8]。

联系发现研究并不是针对两个实体之间的简单联系,而是由多个联系组合而成的复杂联系。例如:"文献A引用文献B"是一个很普遍的联系,"文献B引用文献A"也是一个很普遍的联系,但是可以明显看出,两者的复合"文献A引用文献B,文献B也引用了文献A"是一个值得关注的异常联系[2]。联系发现系统是要自动地识别和推断出诸如此类的异常复合联系。在实际应用中联系发现系统是同时考察多个实体及相互之间的复合联系,而不是单个孤立的实体或简单的联系。

与传统的数据挖掘方法相比,联系发现具有5方面的特殊之处:(1)数据是异构的,来自多个来源,包括:人、事件、对象、动作、计划和组织。(2)在联系发现中,节点表示实体而联系是它们之间的关系;在传统的数据挖掘方法中,节点表示变量而联系表示变量之间关系的概率。(3)联系发现要判定基于特殊图论结构的数据实例和值得关注的模式之间匹配的概率。(4)所有联系发现问题都是通过抽样的数据来评估整个群体,而样本数量一般都很少;(5)数据价值会随着时间流失,所以联系发现的中心问题是选择何时做出决定[9]。

同时,联系发现也面临着不少难以解决的难题:(1)数据范围广泛,从毫无组织结构的资源(报道、新闻故事等)到有高度组织结构的资源(传统的关系数据库),其中的无组织结构资源需要先进行预处理。(2)数据是复杂的、多关系的且包含许多不相关的连接(连接灾难(Curse of Link))。(3)数据是有噪声的、不完备的、有错误的或有多个别称。(4)数据的来源是异构的、分布的和巨量的[10]。

本文对目前联系发现的研究现状进行综述并展望未来的发展趋势。本文的组织如下:在第1节,综述了联系发现的概念、范围、特点及难点等;在第2节,详细介绍联系发现的几种

*)本课题得到国家自然科学基金(No.60503022)的资助。陈 飞 博士生导师,研究领域:人工智能、机器学习。商 琳 硕士研究生,研究领域:机器学习、数据挖掘、神经网络;陈世福 教授,

主要方法：无监督的联系发现方法（新颖联系发现）、使用归纳逻辑程序技术挖掘关联数据的联系发现方法、多假设反演推理的联系发现方法、基于相关分析的联系发现方法以及 KOJAK 组队探测器；在第 3、4 节，分别讨论了联系发现系统的性能评估方法与联系发现的置信区间度量方法；在第 5 节，简要描述联系发现的一个具体应用项目实例——EELD；最后探讨了目前联系发现研究中存在的问题和未来发展趋势。

2 联系发现的几种主要方法

2.1 无监督的联系发现方法

在这里，挖掘的数据集合是一组用二元关系连接的实体集合，也就是说，数据集合中的每个对象都是独立的个体，而个体之间由各种二元关系连接，即数据集就可以表示成为网络的形式。同时假设数据提供丰富的关系词汇且不同类型的连接表示不同的语义关系，基于此假设，Lin 等提出了无监督联系探索方法，又称为新颖联系发现（Novel Link Discovery）[2]。

2.1.1 新颖路径探索[2]

新颖路径问题定义如下：在网络中，针对任意一对实体间存在的多个连接路径，寻找到最值得关注的路径。因为在大多数情况下稀有的路径可能就是值得关注的路径，所以往往需要先寻找稀有路径，然后考察该稀有路径是否是值得关注的。

每一条路径都可以写成如下的形式：

$$e_0 \xrightarrow{r_0} e_1 \xrightarrow{r_1} e_2, \cdots\cdots, \xrightarrow{r_{n-1}} e_n$$

这里 $e_i(i=0,1,2,\cdots,n)$ 是实体，$r_i(i-0,1,2,\cdots,n-1)$ 是相连关系，定义 e_0 为源节点，e_n 为目标节点。路径的类型就是关系序列$[r_0,\cdots,r_{n-1}]$。

式(1)所示的是路径稀有度的定义。

$$rarity(p)=1/N(p) \tag{1}$$

其中 $N(p)$ 是与 p 相似的路径数目。

根据节点类型的不同，定义出 4 类相似路径，分别求出每种类型中的相似路径数目：

N1：类型相同、节点和目标节点都相同的路径数目；

N2：类型相同、源节点相同的路径数目；

N3：类型相、目标节点相同的路径数目；

N4：类型相同的路径数目。

这样，只要求出稀有度最高的路径，就可能得到最值得关注的路径。该方法直接从路径本身考察，而不会被联系的字面含义所误导。

2.1.2 新颖回路发现[2]

与上述路径表示相似，每一条回路都可以写成如下的形式：

$$e_0 \xrightarrow{r_0} e_1 \xrightarrow{r_1} e_2, \cdots, \cdots, \xrightarrow{r_{n-1}} e_0$$

即其源节点和目标节点是同一点，因此 N1、N2、N3 是相同的值，回路稀有度的度量只有两类情况：1/N1，1/N4。具体思路与 2.1.1 节中稀有路径求法相似。

2.1.3 重要节点发现[2]

重要节点探索是对于给定节点，找出连接它的最重要实体。重要的连接不仅取决于路径数目而且取决于路径质量，于是引入联结重要度来衡量路径的质量。

节点 A 和 B 之间的连接重要度，计算公式如式(2)所示。

$$node\text{-}signifcance(A,B)=\sum_{\substack{P_i\in paths\\ between(A,B)}} path\_rarity(P_i) \tag{2}$$

式中 P_i 的 i 用来区别 2.1.1 节中定义的四种路径。

最重要的连接到给定节点 A 的节点求法如式(3)所示。

$$\underset{X}{\operatorname{argmax}}\Big(\sum_{\substack{P_i\in paths\\ between(A,X)}} path\_rarity(P_i)\Big)\underset{X}{\operatorname{argmax}}\Big(\sum_{\substack{P_i\in paths\\ between(A,X)}} \frac{1}{N2(P_i)}\Big) \tag{3}$$

2.1.4 新颖节点发现[11]

新颖节点发现的目标是找出与给定实体之间连接最值得关注的实体。与重要节点发现不同的是，目标不是寻找稀有连接，而是寻找异常连接，这里认为异常的连接便是最值得关注的。异常性是一个相对的度量，取决于其他节点与源节点之间的路径类型不同的贡献。设有路径 p，从源节点 S 到任意目标节点 T，则路径类型 t_p 的贡献为 $N1/N2$（这里 $N1$、$N2$ 的定义见 2.1.1 节）。可以明显看出，贡献 $N1/N2$ 是一个 0 到 1 之间的值。贡献概率的意义是“如果我们选择一条从源节点 S 出发类型为 t_p 的路径，那么它能到达目标 T 的机会”。如果异常性的贡献越大就越表明：对于 S，到达 T 的关系比到达其他节点的关系更加值得关注。

在多关联数据库中探索新颖节点的无监督方法分为 5 步：

①对于任意目标节点 T，列举所有 T 和源节点 S 之间的路径类型 t_p。

②对于每一类 t_p，求出贡献 $N1/N2$。

③把每一种路径类型看作 T 的一个特征，每个特征值为该类型的贡献。如果 T 和 S 源节点之间的路径中不存在某路径类型，则该类型的特征值为 0。

④对于网络中每一个可能的目标节点，重复步 1～步 3。

⑤对于一个有 m 个节点和 $n-1$ 种路径类型的网络，最终可求出 $m-1$ 个 n 维的点。利用 k 近邻算法求出最异常点，该最异常点就是对于源节点做出异常贡献的新颖节点。

2.2 使用归纳逻辑程序技术挖掘关联数据的联系发现方法[3]

大多数数据挖掘方法针对的数据是一种特征向量形式的数据，不能解决多关系数据问题，为此可引入归纳逻辑程序设计来处理多关系数据。归纳逻辑程序设计（Inductive Logic Programming，以下简称 ILP）是用一阶谓词逻辑表示的数据和规则学习的方法。ILP 是关联规则归纳研究使用的最主要方法之一[12,13]。关联数据库很容易转换成一阶谓词逻辑，可以作为 ILP 的数据资源。ILP 的目标是从数据库推断出规则，该类数据库必须已给出背景知识和其它关系的逻辑定义[14]。

ILP 问题的正式定义如下。在一阶谓词演算中 if-then 规则被正式称为 Horn 子句。

已知：

背景知识 B，一组 Horn 子句的集合；

肯定的实例 P，一组 Horn 子句的集合（类型背景文字 typically ground literals）；

否定的实例 N，一组 Horn 子句的集合（类型背景文字 typically ground literals）；

寻找：假设 H，一组 Horn 子句的集合，满足：

$-\forall p\in P: H\cup B\subset p(completeness)$

$-\forall n\in N: H\cup B\not\subset(consistency)$

目前已经提出多种解决 ILP 问题的算法[15]，并且这些算法已广泛引入到了数据挖掘之中[16]。Mooney 等人[3]把该方法应用于关联数据挖掘中解决联系探索问题，在多个证据抽取和联系发现研究计划（EELD）数据集合上的实验表明该方

法在联系发现方面取得的效果很好。因此，有人称基于逻辑的方法为“下一代”数据挖掘系统的重要主题之一。

2.3 多假设反演推理的联系发现方法[17]

在 CADRE 系统（Continuous Analysis and Discovery from Relational Evidence）中使用反演推理方法来为观察到的事实给出一个最佳的解释，实现过程是首先产生假设，接着对假设进行评估。该方法的理论根源是用基于过程和规划的方法对自然语言的理解进行平面图识别[18,19]，其中的关键是一个有效的假设产生机制而不是对已产生假设的评估。

为处理相关数据的不完整和稀疏性，采用一种基于一定限制的分层模式表示方法。这样做的好处有两方面：(1)可以增量地从不同层次上填充已有假设，(2)允许对残缺数据进行基于限制的推理。采用的假设产生方法有两种：自底向上的规则触发推理法（triggering rules）和自顶向下的反演假设提炼法（abductive hypothesis refinement），把两者结合使用，可以有效地从庞大的假设空间中产生有限个假设。首先通过假设产生器产生一组局部假设，每个假设涉及一个潜在的异常事件（威胁）；接着结合反演推理和概率模型，假设评估器从产生的局部假设集合中产生一个全局假设。

该方法有两大优点：(1)特别适用于数据是海量的而目标却是稀疏和不完全的情况；(2)产生规则的方式是增量式的，便于扩充。

2.4 基于相关分析的联系发现方法

Zhang 等提出了一种基于相关分析的联系发现方法（Link Discovery based on Correlation Analysis，简称 LDCA）[10]。LDCA 用相关性度量的方法来考察两个数据项之间的模式相似度，并以此推断两者联系的强弱。LDCA 分为三个基本步骤：(1)联系假设，定义任意两项之间存在一个相关性度量函数，函数值表示相关性大小，其范围为[0,1]；(2)联系产生，求出任意两项的相关性度量函数值，即求出任意两项的相关性大小，再把结果表示为一个加权的多边完全图；(3)联系确认，定义一个新的函数，把这个完全图匹配到它的某个子图，该子图就是一个“产生团体”，该团体内各项之间联系紧密。

LDCA 应用于洗钱犯罪调查时，在联系假设阶段中，先采用了聚类的方法来挑选出两个体之间有价值的金融事务交流，除去不必要的噪声。假设数据库中总共有 n 个实体，则问题可以转化为在 $n+2$ 维欧几里得空间中的聚类问题（n 个实体维、1 个时间维和 1 个事务维）。因为所有实体共享一个时间维，所以也可为每个实体建立一个经济活动的时间直方图，这样聚类处理简化为在（复合）直方图上的分割问题。

在联系产生阶段中，引入基于分层结构的方法（hierarchical composition based，简称 HCB）考察两实体之间的相关性，分为某段时间内的局部相关性和所有局部相关性结合产生的全局相关性。

求出每一对实体间的相关性后，便可得到一个完全图 $G(U,E)$，其中 U 是实体集合，E 是实体间所有相关性的集合，图 G 中各边的权值是两个实体间的相关性。定义阈值 T，和函数 P 表示基于 T 的子图，找出实体间的潜在联系。

2.5 KOJAK 组队探测器

为了从不完整的和有噪声的现实证据中，找出组织或者实体间隐藏的关系，Adibi 等设计出 KOJAK 组队探测器[20]。它的目标是从大型证据数据库中找出隐藏的组队及其成员。KOJAK 系统输入首要和次要证据（存储在关联数据库中），输出产生的组队假设（例如：组队成员列表）。工作过程分为四个阶段：(1)基于逻辑的组队种子产生器，分析首要证据，输出一批组队种子；(2)基于信息论的相互信息模型[21]，为每组找出新的可能候选成员；(3)使用相互信息模型为这些可能的成员排序，顺序是按照它们和种子成员的联系强度；(4)用阈值来裁剪排序后的扩展组队，产生最终的输出结果。KOJAK 组队探测器可以处理异类的、结构不同的证据，解决“连接灾难”问题（Curse of Link），处理噪声，适用于现实中海量的数据库。

KOJAK 组队探测器只是整个 KOJAK 混合联系发现系统的一个模块。KOJAK 系统结合了知识表示和推理的统计聚类技术与数据挖掘领域的分析技术，其中包括多种联系发现部分，例如：模式探测器可用来识别证据中重要的复杂模式，而连接探测器则用来识别重要的和异常的实体及其连接。

3 联系发现系统的性能评估[22]

参照传统分类系统的性能评估方法[23,24]，制定出一系列针对联系发现系统的性能评估准则。

进行性能评估的前提假设是：不考虑真实分类如何产生；真实分类在性能评估中是可用的；联系发现系统和真实分类编制者都可以使用本体论；在系统分类中存在各类断言，这些断言是真实分类和联系发现系统的协议。

监督分类和联系发现系统相似之处在于，目标都是找出给定的数据项是否属于已知的类型。如果把联系发现系统映射成无监督分类问题，则忽略了联系发现系统评估中关键模式类型信息是否可用这个问题。因此，应该把联系发现系统评估问题重定义为一个更广义范畴的监督分类问题：

已知：(1)系统分类的系统数据项：

$\{(sd1,S1),(sd2,S2),\cdots,(sdm,Sm)\}$

(2)真实分类的真实数据项：

$\{(td1,T1),(td2,T2),\cdots,(tdn,Tn)\}$

目标：求出两种分类的相似度。

在监督分类系统性能考察中，最常用的两个度量准则是查准率（Precision）和查全率（Recall）。在联系发现系统性能评估中，需要重新定义查准率和查全率。

相似度的计算函数如式(5)所示。

$$similarity(i,j):\{SD*TD\}\rightarrow[0,1] \tag{5}$$

给定一个系统项，返回与它最相似的真实项的函数如式(6)所示。

$$most-similar-true-item(x) \tag{6}$$

给定一个真实项，返回与它最相似的系统项的函数如式(7)所示。

$$most-similar-systems-item(x) \tag{7}$$

在联系发现系统性能评估中，查准率和查全率的计算方法是式(8)和式(9)。

$$Precision=\sum_i similarity(sdi,tdj)*equal(Si,Tj)/toi=most-similar-true-item(sdi) \tag{8}$$

$$Recall=\sum_i similarity(tdi,sdj)*equal(Ti,Sj)(sdj=most-similarity-systems-item(tdj)) \tag{9}$$

不同的分类错误导致的影响是不同的，用重定义的查准率和查全率来计算影响代价，基于此衡量联系发现系统的性能。这里的真实分类可以由该领域的专家或者自动化的 Agent 生成[25]。

4 联系发现的置信区间度量

目前，大多联系发现的算法是确定的，构造的假设是没有概率限定的。但在所有实际应用中，可用数据都是从数据群体中抽样的，并没有准确地反映整个群体的概率性质。发现的知识和蕴含的假设实际上有概率的，其不确定性是需要度量；需要考察联系发现算法的信任度问题。Adibi 等人提出了采用引导重抽样方法(bootstrap resampling method)[26]来度量这些假设的置信区间[5]。具体步骤如下：

(1) 对于数量为 n 的事件样本，用随机"引导"抽样方法产生替代样本。

(2) 用"成组探测"的方法把抽样事件表示成一幅图，扩充该图并找出个体之间的关系强度和所有的 $g_i \in G$(g_i 表示第 i 个组队)。

(3) 对于每个 g_i，列出所有高于阈值的交互信息 p_i。在这张列表中保存 $L^b(g_j)$，$b=1,\cdots,B$。$L^b(g_j)$ 的每一行表示：在第 b 次运行中，某个体和 g_j 之间的交互信息。

(4) B 取足够大，算法循环 B 次，重复步骤 1，2，3。

(5) 估计每个值得关注的参数的"引导标准错误"。从原来样本统计量中减去平均的引导统计量，就可以得到值得关注的统计偏置评估。

(6) 通过一些基本的数学转化，就可以求出置信区间。

5 联系发现的应用——EELD[27]

作为一个较新的数据挖掘研究领域，联系发现已被广泛应用到很多社会领域之中。目前，比较引人关注的联系发现应用项目是证据抽取和联系发现研究计划(EELD)。"9.11事件"发生之后，随着如何侦查和预防恐怖事件成为社会焦点，美国国防部高级研究项目机构(DARPA)组织了证据抽取和联系发现研究计划(EELD)，把联系发现应用到反恐之中[28]。EELD 的目标是开发出可从海量复杂数据资源中自动发现、抽取并结合稀少证据的技术和工具。使用这些技术和工具来提高侦察能力，及早发现出危害国家安全的不对称威胁，在这些松散的小组织还没有开始袭击之前挖掘出他们活动的蛛丝马迹。EELD 最终要实现的是从多类别的实体和多类别的联系中学习模式的能力[29]。

这里以一个简单的洗钱调查为例，说明 EELD 中应用联系发现的方法。"公司 A 是制造商/销售商"、"公司 A 购买设备/产品"和"公司 A 卖出设备/产品"等单独的证据都不能够反映是否存在洗钱犯罪。但是如果我们综合考察各项的组合及各项之间的联系，比如"公司 A 是制造商，A 购买了设备，接着卖出产品"，可以看出这可能是一个正常的商业行为；但是比如"公司 A 是制造商，A 高价购买设备，接着低价卖出设备"，则这可能是一个可疑的异常商业行为。当然，这只是一个简单的实例，具体调查中可能涉及多个公司和多个看似无关的商业事务，从中发现可疑的公司团体和复杂的洗钱方式。通过设计联系发现系统，自动识别和推测出各种不同的可疑异常商业行为，而不是仅仅调查孤立的证据和简单的事务，这样可以有效地进行洗钱证据的调查。

在该项目中，Kovalerchuk 和 Vityaev 把复杂证据相关性和联系发现相结合，研究出一种实用的混合证据相关技术(Hybrid Evidence Correlation，简称 HEC)[30]。HEC 由一阶逻辑(FOL)和概率语义推理(PSI)两部分组成。分为五部分：(1)证据 D 集合，属性集合{A}以及包括两、三个论据的谓词集合{P}组成；(2)一个领域的本体论，起始于{A}终止于{P}；(3)正常模式{N}与可疑模式{S}的形式化定义；(4)模式分类：统计显著和不显著模式，来捕获重要的罕见事情；(5)潜在的可疑模式产生器(假设产生器)G；(6)模式/假设评估器 E；(7)可疑模式选择器 L。

目前，在 EELD 中有多种改进的关联数据挖掘技术用于联系发现。除了 2.2 节提到的 ILP 方法，还有(1)基于图论的关联学习，特别是其中的 SUBDUE[31]已被应用于 EELD；(2)概率关联模型[32]，把一阶谓词和贝叶斯网络两者相结合；(3)关联特征构造[33]，目前已有把基于此方法的 PEOXIMITY 系统[34]应用于联系发现的研究工作。

可以预见，还有很多领域的应用有待探索，联系发现将会在社会生活的很多方面起到巨大的作用，比如法庭取证、金融领域、信息检索领域以及医学领域等。

6 未来主要的研究趋势

今后的研究趋势，主要有以下的四个方向：

(1) 联系发现的研究与联系发现系统的设计。设计联系发现方法时要全面考虑数据挖掘算法的效率、有效性和正确性[35]。针对待研究数据的特点和目前方法存在的问题，提出更好的改进方法。联系发现方法中既要考虑到联系的数量多少，也要考虑到联系的内容是否重要，同时对于特定的应用也需引入一定的专业知识来取得更好的效果。

(2) 引言中已指出联系发现研究还存在许多难题需要解决。虽然已有的部分算法已提出了一些解决方法，但是如何处理不同组织或无组织的数据，如何考虑"连接灾难"与噪声数据，以及如何面对海量的、异构的或是分布的数据等，仍将是联系发现研究中长期存在的挑战。

(3) 虽然已有一些间接验证联系发现的可信度和性能的方法，但目前还没有直接的和准确的判定方法，因此联系发现的可信度问题和联系发现方法的性能评估也是一个需要解决的焦点问题。这方面研究可从特定的应用知识和挖掘算法本身这两方面综合考虑。

(4) 数据挖掘领域已从传统的结构数据挖掘[6]扩展到无结构的数据挖掘[36]、时间序列数据挖掘[37]、文本挖掘[36]、网络挖掘[38,39]等。随着研究的不断深入，联系发现可以应用到更多的数据挖掘领域和社会领域[40]之中。

结束语 联系发现的主要目标是从范围广泛的实体中确认它们之间的联系，这些实体包括物体、事件、人物、组织和计划等。本文主要从四个角度讨论了联系发现：概述了联系发现的几种主要方法；介绍了联系发现的性能评估方法和置信区间度量的方法；联系发现的具体应用，描述了证据抽取和联系发现研究计划(EELD)；分析了目前联系发现研究中存在的难点和未来的研究展望。可以预见联系发现将在数据挖掘中发挥越来越重要的作用，并将被更广泛地应用到社会生活的各个方面之中。

参考文献

1 Lise, Getoor. Link Mining: A new Data Mining Challenge. ACM SIGKDD Explorations Newsletter, 2003, 5(1): 84～89

2 Lin Shou-de, Chalupsky H. Unsupervised Link Discovery in Multi-relational Data via Rarity Analysis. In: Proceedings of the Third IEEE International Conference on Data Mining. Melbourne, Florida, USA, 2003

3 Mooney R J, Melville P, Rupert Tang L P, et al. Relational Data Mining with Inductive Logic Programming for Link Discovery. In: Proceedings of the National Science Foundation Workshop on Next Generation Data Mining. Baltimore, Maryland, USA, 2002

4 Senator T. Evidence Extraction and Link Discovery Program. DARPATech 2002, 2002

5 Adibi J, Cohen P R, Morrison C T. Measuring Confidence Inter-

vals in Link Discovery: A Bootstrap Approach. ACM SIGKDD 2004. Seattle, Washington, USA, 2004
6 Han J, Kamber M. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann, 2000
7 Goldberg H, Senator T. Restructuring Databases for Knowledge Discovery by Consolidation and Link Formation. In: Proceedings of the First International Conference on knowledge Discovery and Data Mining, Menlo Park. CA: AAAI press, 1995
8 Adibi J, Chalupsky H, Grobelnik M, et al. KDD-2004 Workshop Report Link Analysis and Group Detection (LinkKDD-2004). SIGKDD Explorations, 2004, 6(2): 136～139
9 Adibi J. Real world characteristics of link discovery databases. Workshop on Data Mining for Counter Terrorism and Security with the Third SIAM International Conference on Data Mining (SDM 2003), 2003
10 Zhang Zhongfei(Mark), Salerno J J, Yu Philip S. Applying Data Mining in investigating Money Laundering Crimes. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Washington D. C., 2003. 747～752
11 Lin Shou-de, Chalupsky H. Using Unsupervised Link Discovery Methods to Find Interesting Facts and Connections in a Bibliography Dataset. ACM SIGKDD Explorations Newsletter, 2003, 5(2): 173～178
12 Lavrac N, Dzeroski S. Inductive Logic Programming. Techniques and Applications. Ellis Horwood, 1994
13 Muggleton S H. Inductive Logic Programming. New York, NY: Academic Press, 1992
14 Muggleton S H. Inductive Logic Programming. New York, NY: Academic Press, 1992
15 Dzeroski S, Lavrac N. An introduction to inductive logic programming. Relational Data Mining. Springer Verlag, Berlin, 2001
16 Dzeroski S. Relational data mining applications: An overview. Relational Data Mining. Springer Verlag, Berlin, 2001
17 Ploch N J, Hunter D, White J V, et al. Multi-Hypothesis Abductive Reasoning for Link Discovery. ACM SIGKDD-2004. Seattle, WA, USA, 2004
18 Motivation C E. Analysis, Abductive Unification, and Nonmonotonic Equality. Artificial Intelligence, 1988, 34(3): 275～295
19 Schank R C, Abelson R P. Scripts, Plans, Goals, and Understanding. Lawrence Eribaum Associates, Potomac, Maryland, 1977

（下转第 131 页）

（上接第 122 页）

错误和技术人员引入的新错误），课件主程序错误，功能性错误等等。

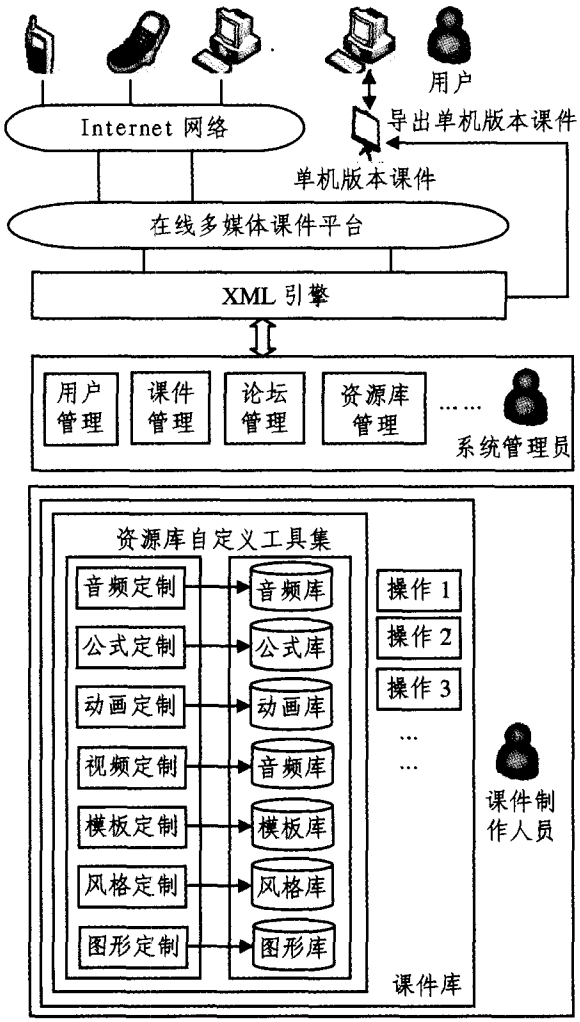

图 4 新型远程教师课件开发平台

由于测试的目标是暴露课件中的错误，从心理学角度来看，由课件开发人员自己进行测试是不恰当的。因此，在综合测试阶段，我们各个课件制作小组交叉进行测试，同时也请远程学院的学生来进行测试。测试任何产品都有两种方法：黑盒测试和白盒测试。不同开发小组的互相测试主要是偏重于白盒测试，学院学生的测试则为黑盒测试。

课件维护是课件开发生命周期的最后一个阶段，课件维护包括改正性维护、适应性维护、完善性维护、预防性维护等等。我们在课件维护过程中不断保存维护记录，并对维护活动进行评价，如每门课件出错的次数、每种错误的花费的总人时数、每门课件维护需要的时间、课件出错给学生学习带来的不便程序等等。

3 课件开发技术展望

课件开发是一个长期的过程，是一个不断学习进步的过程。今后的课件开发将向适时、互动、多元化方向发展[3]。我们展望了一种全新的远程教师课件开发平台，如图 4 所示。这种开发模式同远程课件发布平台相结合，可以做到课件的适时快速开发、更新。只要稍加培训，老师都可以使用这个平台来进行课件制作，同时也可以适时更新课件，解决了以前制作和更新课件只能由制作小组进行的弊端，大大缩短了开发更新周期。在这个平台里集成“傻瓜式”动画制作平台，让普通人员都可制作简单的动画，也可开发出章节标题制作平台，让教师选择某一标题模版后，可进行标题制作。教师制作出的动画和标题等可暂存在“个人制作库中”，在需要的地方加入课件内容。

课件的另一个发展方向是基于网络的游戏型课件[4]，即有明确的教学目标，采用游戏的形式进行教学，教学内容贯穿于整个课件游戏中，在广域网或局域网范围内使用，面向个人自学或课堂教学的交互性多媒体课件。基于网络的游戏型课件较适合培养低年龄段学习者解决实际问题的能力和培养协作意识，也有利于激发学习者的学习兴趣，维持学习者持久的学习动机。

结论 总而言之，课件开发是一个软件开发的过程。采用软件工程理论、技术来开发网络多媒体课件，是课件资源建设的一个方向。另外，课件设计人员和开发人员应当更多地从用户的角度来考虑，更好地实现交互和反馈的功能。

参 考 文 献

1 陈德人，等. 现代远程教育理论与实践探索. 浙江大学出版社，2003
2 丘威，李代平. 基于面向对象软件工程方法的 CAI 开发与实现[J]. 上饶师范学院学报，2003(1)
3 龚玉清，张琴珠. 多媒体课件开发现状的调查研究[J]. 教育传播与技术，2005(42)
4 吴琼，等. 基于网络的游戏型课件在教学中的应用[J]. 中国远程教育，2004(6)

的局限性是生成的系统树状图是很多种模式组合的结果，这使分析员很难决定最终应该选择哪一种。我们提出的算法结合这两种方法的优点，既克服了 ART 神经网络不同的参数可导致不同结果的局限性，又用系统树状图来辅助可视化地选择簇数量。

表 1　本文算法和 K-means 算法的比较

算法 / 结果	K-means 簇数=3	K-means 簇数=2	该算法 簇数=3	该算法 簇数=2
误聚类的模式数/输入模式数	14/186	5/186	2/186	0/186

聚类算法的重要的问题是结合应用领域来聚类数据。因此，我们进一步的工作包括两个方面：一个方面是结合应用领域，使我们的算法应用于大的数据集并聚类任意不规则的数据分布形态，因为使用二次连接的神经网络能够处理任意的数据分布形态，所以聚类任意不规则的数据形态也是可实现的。另一方面是结合不同的应用领域，为算法的参数选择合适的值。

参 考 文 献

1 Krishnapuram R, Kim J. A note on the Gustafson-Kessel and adaptive fuzzy clustering algorithms. IEEE Trans on Fuzzy Systems, 1999, 7(4): 453～461

2 Dave R N. Fuzzy shell-clustering and application to circle detection in digital images. Intern Journal of Genera Systems, 1990, 16: 343～355

3 Dave R N. Use of the adaptive fuzzy clustering algorithm to detect lines in digital images. Intell Robots Comput Vision VIII, 1989, 1192: 600～611

4 Gath I, Geva A B. Unsupervised optima Fuzzy Clustering. IEEE Trans on Pattern Analysis and Machine intelligence, 1989, 11: 773～781

5 Eltoft T, de Figueiredo R J P. A new neural networks for cluster-detection-and-labeling. IEEE Trans. on Neural Networks, 1998, 9(5): 021～1035

6 Geva A B. Hierarchical unsupervised fuzzy clustering. IEEE Trans on Fuzzy Systems, 1999, 7(4): 723～733

7 Su M C, Chang H C. A new model of self-organizing neural networks and its application in data projection. IEEE Trans on Neural Networks, 2001, 112(1): 153～158

8 Su M C, Chou C H. A modified version of the k-means algorithm with a distance based on cluster symmetry. IEEE Trans on Pattern Analysis and Machine Intelligence, 2001, 23(6): 674～680

9 Su M C, Chou C H. A competitive learning algorithm using symmetry. IEEE Trans on Fundamentals of Electronics, Communications and Computer Science, 1999, E82-A(4): 680～687

10 Su M C, Liu T K. Application of neural networks using quadratic junctions in cluster analysis. Neurocomputing, 2001, 37: 165～175

11 Su Mu-Chun, Liu Yi-Chun. A hierarchical approach to ART-like clustering algorithm. Neural Networks. In: 2002. IJCNN '02. Proceedings of the 2002 Internationl Joint Conference, vol1, May 2002. 788～793

12 Eltoft T, deFigueiredo R J P. A new neural networks for cluster-detection and-labeling. IEEE Trans on Neural Networks, 1998, 9(5): 1021～1035

13 DeClaris N, Su M C. A novel class of neural networks with quadratic junction. In: IEEE International Conference on Systems, Man, and Cybernetics, 1991. 1557～1562

14 DeClaris N, Su M C. Introduction to the theory and application of neural networks with quadratic junctions. In: IEEE International Conference on Systems, Man, and Cybernetics, 1992. 1320～1325

15 Carpenter G, Grossberg S. Adaptive resonance theory: stable self-organization of neural recognition codes in response to arbitrary lists of input patterns. In: Proc. 8th Annu Conf Congnitive Sci Soc,. 1986. 45～62

16 Carpenter G, Grossberg S, Rosen D B. Fuzzy ART: Fast Stable learning and Categorization of Analog Pattern by an Adaptive Resonance System. Neural Networks, 1991, 4: 759～771

（上接第 127 页）

20 Adibi J, Chalupsky H, Melz E, et al. The KOJAK Group Finder: Connecting the Dots via Integrated Knowledge-Based and Statistical Reasoning. In: AAAI, 2004. 800～807

21 Adibi J. Link Discovery via a Mutual Information Model: From Graphs to Ordered Lists. In: DIMACS Workshop on Applications of Order Theory to Homeland Defense and Computer Security, Rutgers, 2004

22 Upal M A. Performance Evaluation Metrics for Link Discovery Systems. In: Proceedings of the Third International Conference on Intelligent Systems Design & Applications, Springer Verlag, New York, 2003. 273～282

23 Weiss S, Kulikowski C. Computer Systems That Learn Classification and Predication Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. San Mateo, CA: Morgan Kaufmann, 1991

24 Upal M A, Neufeld E. Comparison of nonhierarchical Unsupervised Classifiers. In: Proceedings of the International Conference on Information, Statistics and Induction in Science, Singapore: World Scientific, 1996

25 IET. Performance Evaluation Specifications for EELD: [Technical Report]. Information Extraction & Transport Inc. 2002 (http://www.iet.com/Projects/EELD)

26 Davision A C, Hinkley D V. Bootstrap Methods and their Application. Boston: Cambridge University, 1997

27 http://infowar.net/tia/www.darpa.mil/iao/EELD.htm, 2005-4-12

28 MckKay S J, Woessner P N, Roule T J. Evidence extraction and link discovery (EELD) seedling project, database schema description. (version 1.0): [Technical Report]. 2862. Veridian Systems Division, 2001

29 http://www.rl.af.mil/tech/programs/eeld/, 2005-4-12

30 Kovalerchuk B, Vityaev E. Correlation of complex evidences and link discovery. In: The Fifth International Conference on Forensic Statistics, Venice, 2002

31 Cook D J, Holder L B. Graph-based data mining. IEEE Intelligent Systems, 2000, 15(2): 32～41

32 Friedman N, Getoor L, Koller D, et al. Learning probabilistic relational models. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence, 1999. 1300～1307

33 Kramer S, Lavrac N, Flach P. Propositionalization approaches to relational data mining. Relational Data Mining. Berlin: Springer Verlag, 2001. 262～291

34 Neville J, Jensen D. Iterative classification in relational data. In: Papers from the AAAI-00 Workshop on Learning Statistical Models from Relational Data, Austin, TX: AAAI Press/ The MIT Press, 2000

35 Zhou Zhi-hua. Three perspectives of data mining. Artificial Intelligence, 2003, 143: 139～146

36 Mack R, Hehenberger M. Txt-based knowledge discovery: search and mining of life-sciences documents. Drug Discovery Today, 2002, 7(11): 89～98

37 Hosking J R M, Pednault E P D, Sudan M. A statistical perspective on data mining. Future Generation Computer Systems, 1997, 13: 117～134

38 Aggarwal C C, AL-Garawi F, Yu P S. Intelligent crawling on the World Wide Web with arbitrary predicates. In: Proc. ACM WWW, 2001

39 Toyoda M, Kitsuregawa M. Creating a Web community chart for navigating related communities. In: Proc. ACM HT, 2001

40 Kovalerchuk B, Vityaev E, Ruiz J F. Consistent and Complete Data and "Expert" Mining in Medicine. In: Medical Data Mining and Knowledge Discovery, Springer, 2001. 238～280