

KDD 中知识的自动评价系统^{*}

秦艳霞 杨炳儒

(北京控制工程研究所软件中心 北京100080)

An Automatic System for Evaluating the Discovered Knowledge in KDD

QI Yan-Xia YANG Bing-Ru

(Dept. of Computer Science and Engineering, Beijing University of Science and Technology, Beijing 100083)

Abstract How to automatically evaluate the discovered knowledge is very important in KDD, but the research on this problem is very little. The paper gives one automatic system for evaluating the knowledge, and provides many solutions. First some relative concepts are described, and the construction of this system is given, and uses the case to prove it.

Keywords Knowledge discovery, Knowledge evaluation, Data mining

1. 引言

在知识发现过程中,通过挖掘算法产生大量的模式,例如斯坦福大学对一个有3万条记录的调查数据库中使用关联规则的发现算法,产生了2万多条的规则^[1],而大多数规则是显而易见和毫无意义的。当这么多规则呈现在用户面前时,用户会无所适从而无法把精力集中在他们真正感兴趣的知識上。如何对它们进行评价,选取出用户感兴趣的和有用的知识成为至关重要的一环,故对知识的自动评价的研究具有重要意义。然而,在当今 KDD 领域以数据挖掘算法为中心的主流发展中,知识评价问题的研究很少有人问津。

数据库上的知识发现(KDD)的结构多种多样,从与知识评价的结合方式的角度来划分,主要有以下四种形式:(见图1)

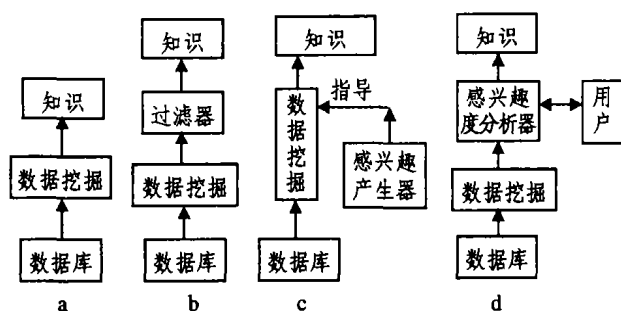


图1 KDD 的结构形式

图 a 是目前典型的知识发现方法^[2~5],它没有知识评价过程,数据挖掘的结果就是最终输出的知识。导致用户面对如此众多的知识,无法把精力集中在他们感兴趣的知識上。

图 b 中的知识发现系统相对图 a 的要实用^[6],它在数据挖掘后面加了一个过滤器,用来筛除掉那些不符合给定评价标准的知识。但是缺少背景知识和与用户的交互过程,而这些恰恰是很关键的,因为知识发现是面向具体的应用并且最终使用者是用户。另外,它仅有一个过滤器,只能删除掉某一类型的知识,而这在实际应用中是远远不够的。

图 c 这种形式从理论上讲是最具吸引力的^[6],因为数据

挖掘是在感兴趣产生器的指导下进行的,只挖掘那些用户感兴趣的知識,从而使搜索空间减小,挖掘速度加快。它不会产生用户不感兴趣的和无用的知識,提高了系统的效率。感兴趣产生器中最重要的组成部分是背景知識和用户兴趣方面的知識。要想使这个系统有效地运行,必须使输入的这些知識是完全的,不然会丢掉许多有用的知識。这便遇到人工智能领域主要的一个难题,即知識获取问题。并且用户的兴趣是不断变化的,不同时刻感兴趣的方面和对同一问题的感兴趣程度是不同的。因此要确定在特定时刻用户感兴趣度是几乎不可能的。同时,当不同的用户使用同一个知識发现系统时,感兴趣产生器要随着改变,这便需要重新运行数据挖掘过程;但是数据挖掘是针对大型数据库进行操作的,运算量很大,这便不利于对于同一个数据库不同用户的使用。再者,对同一个用户而言,使用多少次这个系统,就需要进行多少次数据挖掘过程,降低了系统的使用价值。

图 d 是图 b 的一种扩展^[7],其中感兴趣分析器是由一系列感兴趣的度量单元组成,它们是并行操作的,分别对不同类型的知識、同一类型的不同形式进行评价。通过不断的人机交互,一方面用户可以输入有关的背景知識,另一方面通过这一阶段知識评价的结果可以激发用户新的兴趣或对原有的兴趣进行实时的修改。再者,对于不同的用户,他们使用的是同一个数据挖掘的结果,不同的是各个度量单元,改变时不需要直接对数据库操作。与图 c 相比,系统的运算量大大减少。当然,这种方式的不足是要保证不同度量单元的相互独立,因为最后输出的不仅是知識的表示形式,还有感兴趣度。这样各个度量结果的相互影响成为要考虑的问题。

从上面的分析可以看出:尽管多数的知識发现系统还没有考虑评价过程,但作为知識发现的一项重要内容应越来越受到重视。图 b、c、d 中的系统都加入了评价这一个环节,方式各不一样,都各有其优缺点。当获取的背景知識很少时应该使用图 b 特别是图 d 中的知識发现方法。这样既可以避免丢失知識,又有利于背景知識的进一步获取。而当领域知識比较完全,同时用户的兴趣相对稳定时,可考虑使用图 c 中的知識发现方法。

^{*} 国家自然科学基金重点项目(69835001)资助、教育部科技重点项目([2000]175)资助;北京自然科学基金项目(4022008)资助。

综上所述,目前所有评价方法的本质缺陷在于:仅在 KDD 封闭的系统内部运行,没有形成评价系统,没有形成评价的自主性,没有提升到认知系统复杂性的高度加以研究。

根据人类认知的特点,对事物或知识的评价是一个分层次、逐步求精、客观和主观因素相结合的综合过程。基于此,我们提出了 KDD 中的分层次的自动的知识评价系统。整个评价过程是分层次的:第一层是基本的,是客观地从数据的层面进行的;第二层是有所侧重,强调人机交互和领域知识的参与,是主观的、从规则的层面进行的;第三层是上面两个层次的综合,即综合了主客观两个方面的因素,形成了综合的评价指标。该评价模型与方法基于双库协同机制^[3]的内在机理,将评价过程与基础知识库联系起来,通过计算机实现了人类进行评价的基本过程,符合人类的认知规律,使评价过程具有了人的智能。下面将以关联规则为例,介绍这一评价系统的构造过程,并通过实例给予说明。

2. 客观评价指标(第一层次)

常用的衡量关联规则客观性的指标有两个:支持度和可信度。支持度度量的是规则的实用性(即指发现的知识有实际效用);可信度度量的是规则的有效性(即指发现的模式对于新的数据仍保持一定的可信度)。强规则的定义就是规则的支持度和可信度分别大于最小支持度阈值和最小可信度阈值的规则。Apriori 算法的目的就是找到这些强规则,实际上这种挖掘算法主要考虑的是规则的有效性。

但是强关联规则并不一定是用户感兴趣的,支持度-可信度框架可能误导。事实上,当 A 的出现并不蕴涵 B 的出现时,用这种框架也可能识别出规则 $A \Rightarrow B$ 是有趣的。因此在客观感兴趣度量中,还需要考虑项集之间的相关性。如果关联规则的前件和后件是不相关的,即使规则的支持度和可信度很高也不能被用户接受为感兴趣的关联规则。所以,要对挖掘所得的规则首先进行相关性检验,只有那些前件和后件是强的正相关关系的规则才被选择进行以后的处理。

定义1 A 和 B 的出现之间的相关性度量 corrR 的计算公式为:

$$\text{corrR}_{A,B} = \frac{P(A \cup B)}{P(A)P(B)}$$

如果值小于1,则 A 的出现和 B 的出现是负相关的。如果值大于1,则 A 和 B 是正相关的,意味着每一个的出现都蕴涵着另一个的出现。如果值等于1,则 A 和 B 是独立的,它们之间没有相关性。因此选择相关强度作为关联规则评价的一个客观性指标,它同时反映了规则的前件和后件的相关程度。所以客观的评价指标有三个:支持度、可信度和相关强度。只有满足这三个基本指标的关联规则才可能进行下一个层次的主观评价过程。

3. 主观评价指标(第二层次)

主观评价指标主要体现用户和领域知识的参与等主观因素,是从规则的层面来考虑的,不必考虑数据库中的数据。特点是针对具体的主观评价指标,规则的评价有所侧重。根据 KDD 的定义,与主观评价有关的指标有:新颖性(指发现的模式是新的)、用户感兴趣性(可体现潜在有用性)和简洁性(体现最终可理解性)。相应的度量标准分别为新颖度、用户感兴趣度和简洁度,下面分别给予说明。

3.1 新颖度

新颖度分别表现在发现的规则与基础知识库(主要存放专家输入的领域知识和用户已知的一些规则)中已有规则各项的差异上,分别表现在前件各项的差异和后件各项的差异上(分别从语言变量和同一语言变量的不同语言值的角度),同时前件和后件也分别看待之。

经过第一层次的评价过程的筛选后,新发现的规则数已经大大地减少了。设现在剩余的规则组成的集合为 E,其规则数为 $|E|$;基础知识库中的规则集合记为 K,其规则个数为 $|K|$ 。设 W_i 为 E 中的规则 E_i 相对于 K 的新颖度, $w_{(i,j)}$ 是规则 E_i 与基础知识库中的规则 K_j 之间的新颖度即差别程度。计算 W_i 的图示如下(见图2):

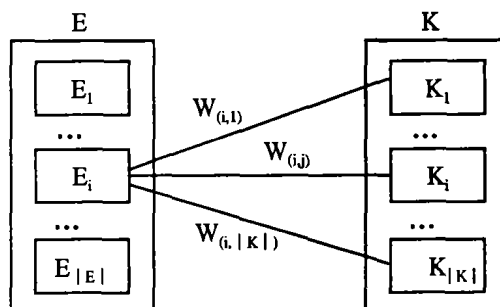


图2 计算 W_i 的示意图

$w_{(i,j)}$ 包含两部分,前件的新颖度 $L_{(i,j)}$ 和后件的新颖度 $Z_{(i,j)}$

定义2 设基础知识库中的规则 K_j 中所有前件所属的语言变量组成的集合为 J,并且 E 中规则 E_i 所有前件所属的语言变量组成的集合为 I。

对 I 中的任一项 I_k ,记 $V_{(i,j)k}$ 为这一项所对应的规则前件与规则 K_j 的前件差异程度,则有

$$V_{(i,j)k} = \begin{cases} 2, I_k \notin J \\ 1 + \text{neg}_k, I_k \in J \end{cases} \quad (1)$$

neg_k 为 I 中的第 k 项的语言值与 J 中同一语言变量对应的语言值之间的差异程度。在它前面加1是为了避免当所有项都在 J 中出现并且对应语言值均相同时,会出现0的情形。这样,前件的新颖度等于前件中各项的差异程度的累加和,即

$$L_{(i,j)} = \sum_{k=1}^{|I|} V_{(i,j)k} \quad (2)$$

经过规则简约后,基础知识库中所有规则的后件的项数均为1,同时通过数据挖掘算法得到的规则后件的项数也为1。所以新发现的规则 E_i 与知识库中的任一条规则 K_j 在后件上有且仅有下面两种可能的关系:

第一种情况:两条规则的后件属于同一个语言变量,这时首先计算两者的语言值对应的矛盾度,则 $Z_{(i,j)} = 1 + \text{neg}$;加1的目的是为了避免当后件是同一个语言变量的相同语言值时出现0的现象。

第二种情况:两条规则的后件不属于同一个语言变量,这时令度量后件差异的数值记为2,即 $Z_{(i,j)} = 2$

计算 $w_{(i,j)}$ 的公式如下:

$$w_{(i,j)} = \frac{L_{(i,j)} \times Z_{(i,j)}}{\max(|I|, |J|)} \quad (3)$$

定义3 E 中的规则 E_i 相对于 K 的新颖度 W_i 的计算公式为:

$$W_i = \frac{\sum_{j=1}^{|K|} w_{(i,j)}}{|K|} \quad (4)$$

3.2 用户感兴趣度

潜在有用性涉及到对所发现规则的进一步应用问题。可能规则中前件或后件中的一部分属性对于将来的决策和分析事务之间的关系具有重要的意义,能够起到直接或间接的作用。这部分属性可能属于两种情况:第一种是它们在领域中的位置比较重要,对其它属性起着很大影响作用,或者规则本身体现了领域中重要的一种关系,这样它们便会引起发现者的关注;另一种情况是反映了用户进行知识发现时的兴趣取向,对于满足当时情况下所需知识的紧迫程度,对这些属性的感兴趣程度和顺序具有一定的时效性,它们的确定是与现实情况紧密相连的。

因此这两种情况下的潜在有用性的衡量标准将由不同的人按照不同的标准来确定:

针对第一种情况:由领域专家按照各种属性在本领域中的重要程度给出相应的感兴趣值,记为 EXP,它是属于0到1之间的值,最重要的属性置为1,最无关紧要的属性置为0;

针对第二种情况:由最终使用本发现系统的用户根据当时的兴趣和与当时问题的相关性给出感兴趣值,记为 USER。它取值于0和1之间,最感兴趣的置为1,不感兴趣的置为0。

需要指出的是:第一种情况所给出的属性的重要性排序具有一定的稳定性,它反映的是本领域中属性之间的语义关系,是从语义信息的角度来评价规则;而第二种情况给出的属性的感兴趣度具有时效性,它是从实际的实时应用和用户的兴趣出发,反映的是规则对用户实际的应用价值,是从语用信息的角度来评价关联规则。这两种情况下的算术平均值作为衡量规则潜在有用性的度量指标,记为 USI。这个指标的计算是在评价前通过与用户和领域专家交互实现的。

3.3 简洁度

简洁度(记为 CN)是用来衡量关联规则的最终可理解程度的指标。它也表现在两个方面:一方面表现在规则项的个数上,如果规则项数很多将不利于对这条规则的理解。因此,规则的项数是一个衡量规则简洁性的逆向指标,即规则的项数越多,规则的简洁性越差;规则的项数越少,规则的简洁性越好。另一方面表现在规则所包含的抽象层次上,规则包含的项的抽象层次越高,它对数据的解释力越强,也越容易理解;相反规则的项的抽象层次越低,它对数据的解释力越差,因此也不容易理解。

4. 综合评价指标(第三层次)

由上述的两个层次的评价可以看出:客观的评价指标有三个:支持度 s、可信度 c 和相关强度 corrR;主观的评价指标有三个:新颖度 W、用户感兴趣度 USI 和简洁度 CN。它们的取值范围并不相同:支持度和可信度取0到1之间的值;相关强度的值大于1;新颖度、用户感兴趣度和简洁度的取值视具体的应用领域和用户习惯而定。鉴于此,作为第三层次的综合评价指标应是主客观度量指标的有机结合。

定义4 综合评价指标——综合评价度(称为规则的兴趣度 RI)定义为上述几个评价指标的几何加权平均:

$$RI = c^{w_1} \times s^{w_2} \times \text{corrR}^{w_3} \times W^{w_4} \times USI^{w_5} \times \left(\frac{1}{CN}\right)^{w_6}$$

其中, $1 \geq w_i \geq 0, (i=1, \dots, 6), \sum_{i=1}^6 w_i = 1$

其中需要解决的重要问题是各个评价指标的权值的确定,而这与应用的领域是密切相关的,同时与最终用户也有很大的关系。如果应用领域是科学领域那么对有效性要求就比较高,因此赋予有效性度量标准的权值较大,而其它的指标的权值较小。如果在商业领域则注重时效性和可理解性以及有用性,所以对简洁性和有用性的度量标准所赋予的权值较大。而对于主要想满足好奇心和进行探索性查询的用户来讲,新颖性则是主要的评价指标,用户的兴趣度的部分应被赋以较高的权值。

5. 实例说明

目前水稻害虫是影响水稻产量的重要因素,据联合国粮农组织估计,在亚洲因水稻害虫危害造成的稻谷损失高达25%。水稻害虫种类繁多,国内已有记载的有几十种,几乎水稻生长过程中每一阶段、不同部位均有害虫为害,三化螟是其中的一种害虫。本论文收集了有关庐江地区三化螟的数据,就以该螟虫数据库作为应用对象,进行数据挖掘。庐江螟虫害数据库有1926条记录,属性分别为雌虫密度,雄虫密度,最低温度,平均温度,降雨量,日照时数等,这6个属性均为多值型属性,对它们利用语言场和语言值的方法处理以后,再利用 Apriori 算法进行挖掘,将挖掘所得的几条规则置于基础知识库中,见图3。然后以其为基础,对其它的规则进行评价。

基础知识库					
删除全部 删除单条纪录 退出					
序号	条件1	条件2	结果	支持度	可信度
1	雌虫密度 很小		雄虫密度 很小	0.64	0.94
2	平均气温 低		雌虫密度 很小	0.25	0.95
3	最低气温 适中		雌虫密度 很小	0.28	0.75
4	平均气温 低		雄虫密度 很小	0.25	0.95
5	最低气温 适中		雄虫密度 很小	0.3	0.79
6	平均气温 适中		最低气温 适中	0.24	0.7
7	雄虫密度 很小	平均气温 低	雌虫密度 很小	0.25	0.99
8	雌虫密度 很小	平均气温 低	雄虫密度 很小	0.25	0.98
9	雄虫密度 很小	平均气温 适中	雌虫密度 很小	0.21	0.84
10	雌虫密度 很小	平均气温 适中	雄虫密度 很小	0.21	0.94

图3 基础知识库中的规则

利用综合评价的算法对所得的规则进行评价,结果以综合评价指标规则兴趣度的大小顺序显示,如图4所示。综合评价的结果受许多因素的影响,上述结果是在用户感兴趣度(取值范围是0到1)USI=1的情况下得到的。

小结 在 KDD 中形成分层次的、主客观评价相结合的自动评价系统的优点主要有以下几个:

(1)立论于双库协同机制的内在机理中,将评价模型与方
(下转第52页)

法和 Sujun Hua 等人提出的支撑向量机方法都取得了很好的预测结果,可见,氨基酸编码方法和网络结构还有很大的探索空间;综合生物学方面的进展,加入更重要的进化信息,蛋白质二级结构的预测精度将会有新的突破。

参考文献

- Kabsch W, Sander C. Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen Bonded and Geometrical Features. *Biopolymers*, 1983, 22: 2577~2637
- Bernstein F C, et al. The Protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, 1977, 112: 535~542
- Anfinsen C B, et al. *Proc. Natl. Acad. Sci. U. S. A.*, 1961, 47: 1309~1314
- Epstein C J, Goldberger R F, Anfinsen C B. *Cold Spring Harb. Symp. Quant. Biol.*, 1963, 28: 439
- Qian N, Sejnowski T J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, 1988, 202: 865~884
- Matthews B W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, 1975, 405: 442~451
- Rost B, Schneider R, Sander C. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.*, 1993
- Rost B, Sander C. Prediction of Secondary structure at better than 70% accuracy. *J. Mol. Biol.*, 1993, 232: 584~599
- Cuff J A, Barton G J. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Struct. Funct. Genet.*, 1999, 34: 509~519
- Salzberg S, Cost S. *J. Mol. Biol.*, 1992, 227: 371~374
- Hayward S, Collins J F. *Proteins*, 1992, 14: 372~381
- Zhu H, Yoshihara I, Yamamori K. Prediction of Protein Secondary structure by Multi-Modal Neural Networks. *IEEE. Trans.*, 2002, 1: 280~285
- Riis S K. Combining Neural Networks for Protein Secondary Structure Prediction. *Neural Networks, Proceedings.*, 1995, 4: 1744~1748
- Rost B, Sander C. *Proc. Natl. Acad. Sci. USA, Biothysics*, 1993, 90: 7558~7562
- Hua S, Sun Z. A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Machine Approach. *J. Mol. Biol.*, 2001, 308: 397~407
- Rost B. Protein structure prediction in 1D, 2D, and 3D, in *Encyclopedia of Computational Chemistry*, 1998. 2242~2255
- Sander B. R. C. Third Generation Prediction of Secondary structure. in *Protein structure prediction: methods and protocols*, Humana Press, 2000. 71~95
- Chothia C, Lesk A M. *EMBO J.*, 1986, 5: 823~826
- Rost B, Sander C, Schneider R. Evolution and Neural Networks: Protein Secondary Structure Prediction Above 71% Accuracy. In: *Proc. of Twenty-Seventh Annual Hawaii Intl. Conf. of System Sciences*, 1994, 5: 385~394
- Kneller D G, Cohen F E, Langridge R. *J. Mol. Biol.*, 1990, 214: 171~182
- Metfessel B A, Saurugger P N. Pattern Recognition in the Prediction of Protein Structural Class. In: *Proc. of the Twenty-sixth Hawaii Intl. conf. on*, 1993, 1: 679~688
- Dill K A. *Biochemistry*, 1990, 29: 7133~7155
- Deleage G, Roux B. *Protein Eng.*, 1987, 1: 289~294
- Owen L, Liang H H, Matthew B. 2001. Data Representation Influences Secondary Structure Prediction using Artificial Neural Networks. In: *Seventh Australian and New Zealand Intelligent Information Systems Conf.* 2001. 411~416
- Ruggiero C. Peptides Secondary Structure Prediction with Neural Networks: A Criterion for Building Appropriate Learning Sets. *Biomedical Engineering, IEEE Trans. On*, 1993, 40(11): 1114~1121
- Jones D T. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J. Mol. Biol.*, 1999, 292: 195~202

(上接第31页)

法置于有基础知识库参与的开放系统中(不局限于 KDD 自身的封闭系统中);

(2)将主观的和客观的度量指标统一在一个三层次的评价体系中,使它们不再是孤立的分析,而是将它们紧密地结合在一起;

(3)将评价过程自动化,所有的相关计算都由计算机编程

实现,最终以一个统一的综合指标呈现给用户,便于用户比较选择;

(4)在主要是自动完成的前提下,也充分地发挥了用户的作用,并且用户所参与的部分(提供了各个语言变量和语言值的感兴趣度等)具有明确的意义,很容易实现。

(5)综合评价的指标是可扩展的,即可以根据需要添加新的指标,可简单地实现与原有的其它指标的综合。

综合评价						
删除		✓ 存入知识库	✓ 存入基础库			↑ 帮助
序号	条件1	条件2	结果	支持度	可信度	规则感兴趣度
1	最长气温 高		平均气温 高	0.26	0.69	6.6816
2	平均气温 高		最长气温 高	0.26	0.89	5.62368
3	最长气温 适中		平均气温 适中	0.24	0.65	5.264
4	雄虫密度 很小	降雨量 很小	雌虫密度 很小	0.43	0.88	3.968
5	日照时数 多		降雨量 很小	0.24	0.82	3.808
6	雄虫密度 很小	最低气温 适中	雌虫密度 很小	0.26	0.88	3.8055
7	雌虫密度 很小	降雨量 很小	雄虫密度 很小	0.43	0.93	3.6765
8	雌虫密度 很小	最低气温 适中	雄虫密度 很小	0.26	0.93	3.51
9	雄虫密度 很小		雌虫密度 很小	0.64	0.89	3.25
10	平均气温 高		降雨量 很小	0.21	0.71	2.5956

图4 综合评价结果

参考文献

- Padmanabhan B, Tuzhilin A. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision support systems*, 1999, 27: 303~318
- Freitas A A. On rule interestingness measures. *Knowledge-Based Systems*, 1999, 12: 309~315
- Agrawal R, Imielinski T, Swami A. Database mining: a performance perspective. *IEEE Trans. Knowledge and Data Eng.*, 1993, 5(6): 914~925
- Quinlan J R. C4.5: program for machine learning. Morgan Kaufmann, 1992
- Dhar V, Tuzhilin A. Abstract-driven pattern discovery in databases. *IEEE Trans. Knowledge and Data Eng.*, 1993, 5(6): 926~938
- Silberschatz A, Tuzhilin A. On subjective measures of interestingness in knowledge discovery. In: *Proc of the 1st Int'l Conf on Knowledge Discovery and Data Mining*. 275~286
- Liu Bing, Hsu W, et al. Finding interesting patterns using user expectations. *IEEE Trans. Knowledge and Data Eng.*, 1999, 11(6): 817~832
- Yang Bingru. A study on double bases cooperating mechanism in KDD(I). *Engineering Science*, 2002, 4(4): 41~45