二元约束满足问题求解的结点开销模型*

薛瀚宏　蔡庆生

(中国科学技术大学计算机科学与技术系　合肥　230027)

E-mail: s9511s04@sun10.cc.ustc.edu.cn

E-mail: qscai@dawn1.cs.ustc.edu.cn

摘要　提出了在二元约束满足问题中以搜索结点个数为衡量标准的求解开销模型,该模型被应用于随机二元约束满足问题的求解开销相变分析中,并且比较了模型所导出的理论开销和实际中的搜索结点个数、约束检查次数、求解时间 3 种衡量标准的开销之间的相似性.在模型的基础上,探讨了求解启发式减少求解开销的作用,给出了一个新的变量选择启发式.

关键词　约束满足,求解开销,相变,求解启发式.

中图法分类号　TP18

关于约束满足问题 CSP(constraint satisfaction problem)的求解开销及其相变分析目前正在得到广泛的研究.在求解过程中,随着约束强度的变化,其开销曲线上存在着一个尖锐的相变区,这已在大量实验中得到证实.[1~3]随着控制变量的改变,当问题的约束条件较弱时,相对地容易找到解;约束条件较强时,又容易证明问题无解;而最困难的问题就在这两种状态之间,称为相变.状态的尖锐的走势变化区,称为相变区.相变区的峰值点称为相变点.在相变点上,Williams 和 Hogg[4],Prosser[5]预测问题的解数的期望值为 1.Crawford 和 Auton 定义 50%的问题有解而 50%的问题无解的位置为交叉点,并认为交叉点和相变点十分接近.[6]Smith 和 Dyer 定义相变区从 1%的问题无解处到 99%的问题无解处.[7]这些以往的工作主要是以实验的方式来了解问题求解的相变特性,本文提出的模型将从理论上给出求解开销曲线的形状、相变点、交叉点以及相变区,并以实验证实.

相变的分析可以给出对问题复杂度的衡量,而且能够给问题求解以一定的启发.如,Gent 等人在相变分析中提出了问题约束度的概念,并以此得出减小问题约束度的启发式.[8]本文根据对随机二元 CSP 的分析模型,给出一个新的求解启发式——最小值域加最强约束的变量选择启发式,与通常的最小值域启发式[9]相比,能显著地提高求解效率.

1　约束满足问题介绍

一个 CSP 包括一个变量的有限集 $V=\{v_1,\ldots,v_n\}$,一个由每个变量的值域组成的值域集 $D=\{D_1,\ldots,D_n\}$以及一个约束集 C,其中每个约束包含一个 V 的子集$\{v_i,\ldots,v_j\}$和一个约束关系 $R\subseteq D_i\times\ldots\times D_j$.求解 CSP 就是为每个变量在其值域中寻找一个赋值,使得所有约束被满足,或是证实这样的赋值不存在.一个约束被满足时,它所约束的变量的赋值组成的元组在约束关系中.在二元 CSP 中,所有的约束关系都是二元的.

二元 CSP 变量之间的约束关系可以用约束图表示,每个变量对应着一个顶点,每个约束对应着连接表示其约束变量的顶点的边.一对变量之间的二元约束关系可以用一个 $d_1\times d_2$ 的布尔矩阵表示,d_1 和 d_2 分别是两个变量值域的大小,矩阵中的值"1"意味着相应的二元组被约束所允许.

在本文的分析和实验中,使用了随机生成的二元 CSP 集,这样的问题集已被广泛应用于分析测试各种 CSP 求解算法之中.[4,10~12]每个问题集可以用 4 个参数刻画:变量个数 n、变量值域大小 d(假设所有变量的值域相同)、两个变量之间存在约束关系的概率 p 以及约束存在时,被约束变量的取值二元组不满足约束的概率 q.其中 p 称为约束密度,q 称为约束强度.若有随机二元 CSP 集(n,d,p,q),则其约束图中边数的期望值为 $pn(n-1)/2$,约束矩阵中"0"值个数的期望值是 qd^2.

---

* 本文研究得到国家自然科学基金和国家教委博士点基金资助.作者薛瀚宏,1972 年生,硕士生,主要研究领域为约束满足问题,数据库中的知识发现.蔡庆生,1938 年生,教授,博导,主要研究领域为机器学习,知识发现.

本文通讯联系人:薛瀚宏,合肥 230027,中国科学技术大学计算机科学与技术系

本文 1997-07-25 收到原稿,1997-11-12 收到修改稿

© 中国科学院软件研究所 http://www.jos.org.cn

生成随机二元 CSP 集(n,d,p,q),主要是约束图和约束矩阵的生成,可以有以下方式:(1) 约束图边密度的期望值和约束矩阵"0"值密度的期望值分别为 p 和 q;(2) 约束图边密度恰好为 p,约束矩阵"0"值密度恰好为 q,在不能严格相等时取最接近的值.[13]显然后者有较小的方差,实验结果在小的样本上就可以有较高的可信度.本文实验用的问题集,以后者的方式生成,并选取一定的参数,使得上述两个密度值严格地与 p 和 q 分别相等.

2　深度优先搜索求解约束满足问题

目前对 CSP 的求解一般采用深度优先搜索.给定问题,首先进行一致性实施,从变量的值域中删除必定不满足某个或某些约束的值.当删除导致某个变量值域为空时,说明问题分支无解,从而产生回溯;然后采用一定的求解启发式,选择一个尚未赋值的变量和相应的对它的赋值.这样就生成了一个新的 CSP,被赋值的变量值域中仅有一个值,即它的赋值,而其他变量的值域不变.对新生成的 CSP 可以递归求解.

一致性实施主要有两种方法:前向检查[9]和弧一致性维护.[14]在前向检查中,每当一个变量被赋值时,在与其有约束关系的变量值域中删除和该赋值不一致(不满足约束)的值.弧一致性维护则在前向检查的基础上,考虑所有变量之间的一致性.当两个变量之间存在约束时,要求一个变量值域中的所有值都至少和另一个变量值域中的一个值相一致.

衡量 CSP 求解开销的主要因素有 3 个:对赋值对(二元组)进行一致性检查的次数、求解过程中生成的结点个数(或回溯次数)和求解时间.下面的讨论中,在提出了基于结点个数的一般性的深度优先搜索求解开销模型后,重点分析了采用前向检查的算法解决随机二元 CSP 的开销,并且从实验中得出:就开销曲线特性而言,用以上标准来衡量开销和模型的结果是基本一致的.

3　求解开销模型

采用深度优先搜索的算法,可以用搜索树来描述其求解过程.假定求解 CSP 时,变量以某种固定的次序被赋值,设为 $v_1,\ldots,v_n$,其初始值域大小分别为 $d_1,\ldots,d_n$.进行了一致性实施之后,从变量的值域中删除了不满足约束的值,搜索树上相应的分支数就减少了.n 变量 CSP 的搜索树至多有 $n+1$ 层,其中根结点是所有变量都尚未赋值的初始状态.记 l_k 为第 k 层的结点个数,$k=0,\ldots,n$.定义第 k 层的分支系数 $m_k=l_k/l_{k-1}$,$k=1,\ldots,n$,即在变量 $v_1,\ldots,v_{k-1}$ 已赋值并进行了一致性实施之后,变量 v_k 的可能赋值个数的期望值.

图 1 是一个 4 变量 CSP 的搜索树,变量 v_1,v_2,v_3,v_4 的初始值域大小分别是 2,3,3,2.其中的实心结点表示问题的解,阴影结点表示分支无解.

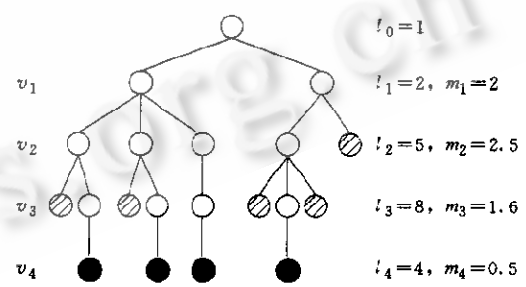

图1　求解CSP的搜索树示例

若已知搜索树各层的分支系数,可以反过来求各层的结点数

$$l_k=\prod_{i=1}^{k} m_i, k=1,\ldots,n. \tag{1}$$

特别地,搜索树最底层的结点个数即是问题的解结点的个数,所以问题的解数

$$S=l_n=\prod_{i=1}^{n} m_i, \tag{2}$$

而搜索树中所有结点的个数

$$T=\sum_{k=1}^{n} l_k=\sum_{k=1}^{n}\prod_{i=1}^{k} m_i. \tag{3}$$

直观上,搜索树中解结点数越多或是总结点数越少,则寻找问题的一个解或是证实问题无解的开销就越小.记求解开销为 C,表示搜索结束时遍历的结点数的期望值.为研究 C,S,T 三者之间的关系,考虑求解过程中未采用启发式的情形,即变量及其赋值以随机方式选取.则可以假定搜索过程是一个随机过程,搜索空间大小为 T,其中以概率密度 λ 出现解结点.不难得到

$$C=T(1-\lambda)^T+\sum_{k=1}^{n} k\lambda(1-\lambda)^{k-1}=[1-(1-\lambda)^T]/\lambda. \tag{4}$$

其求和号内的项是问题有解且第 1 个解结点是搜索树中第 k 个结点时的开销,求和号外的项是问题无解时的开销.令 $\lambda=S/T$ 代入(1)式,则

© 中国科学院软件研究所 http://www.jos.org.cn

$$C=T[1-(1-S/T)^T]/S. \tag{5}$$

并在 T 充分大时有近似式

$$C=T(1-e^{-S})/S. \tag{6}$$

从式中可以看出，当问题接近无解时，开销则接近整个搜索空间的大小，即 $S\to 0$ 时，$C\to T$；而问题有许多解时，开销则只有整个搜索空间的 $1/S$，即 $S\to\infty$ 时，$C\to T/S$.

需要说明的是，模型所得出的只是一种定性的求解开销，可望揭示不同问题的开销之间存在的关系. 从下面对随机二元约束满足问题的求解开销的实验中可以得出，以上所作的假定并不对求解开销的相变特性有显著的影响.

4　随机二元约束满足问题的求解开销分析

4.1　模型上的求解开销

对于给定参数的随机二元 CSP 集 $\langle n,d,p,q\rangle$，考虑求解过程采用前向检查的一致性实施方式，依上述模型，因为第 k 个变量和前 $k-1$ 个变量存在的约束个数的期望值是 $(k-1)p$，而每个约束又只允许了值域中 $1-q$ 比例的值和已有的赋值一致，所以对变量 v_k 可能赋值个数的期望值，即搜索树第 k 层的分支系数有（Ex 表示变量 x 的期望值）

$$Em_k=d(1-q)^{(k-1)p}. \tag{7}$$

下面考虑在各分支系数取其期望值时的求解开销. 由(1)式可得搜索树第 k 层的结点数

$$l'_k=d^k(1-q)^{pk(k-1)/2}. \tag{8}$$

分别由(2)、(3)式得问题的解数

$$S'=d^n(1-q)^{pn(n-1)/2}, \tag{9}$$

搜索空间大小

$$T'=\sum_{k=1}^{n}d^k(1-q)^{pk(k-1)/2}, \tag{10}$$

再由(6)式求得模型上的理论开销

$$C'=T'(1-e^{-S'})/S'. \tag{11}$$

4.2　求解开销的实验对照

由于获得 T 的简单表达式（非级数表示）是困难的，为研究搜索开销随约束强度的变化关系，可以作 q-C 曲线. 图2给出了随机二元 CSP 集 $\langle 8,10,1,q\rangle$ 的模型上的理论开销曲线，并和实验中的实际开销比较. 所使用的算法是：深度优先搜索，前向检查的一致性实施方式，最小值域的变量选择启发式（见后文阐述）和第1个可能值的赋值策略. 算法以 Borland C++ 3.1 实现，在奔腾 100 微机上运行.

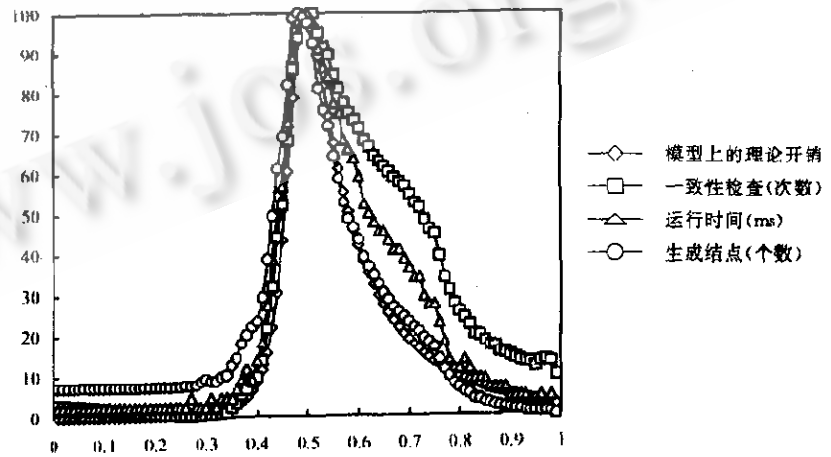

图2　求解问题集 $\langle 8,10,1,q\rangle$ 时模型上的理论开销和实际开销

实验中以约束强度 q 的间隔步长为 0.01 获得数据，每个数据点上的结果都是 100 个问题求解结果的平均值. 为对比曲线的相似性，图中的数据都经过了格式化，作线性变换 $y'=100[y-\min(y)]/[\max(y)-\min(y)]$，其中不同曲

© 中国科学院软件研究所 http://www.jos.org.cn

线取值的对应范围分别是：模型开销 $y\in[1.12,396.32]$，一致性检查次数 $y\in[185.98,3240.40]$，运行时间(ms)$y\in[11.54,35.71]$，生成结点数 $y\in[2.46,79.80]$. 这种格式化是合理的，就运行时间而言，有关输入和输出的开销占用了相当一部分时间，对同一个问题集里的不同问题，这部分开销都大致相同，因而格式化有助于减小额外开销的影响. 从图 2 中看出，虽然衡量问题求解开销的标准不同，但相应的开销曲线的形状是相似的，相变特性是相近的. 也就是说，由模型给出的开销可以作为评价问题之间相对的求解难度的依据.

就$(8,10,1,q)$问题集而言，本模型给出的理论上的相变点位置是 $q=0.50$，不同开销衡量标准的实验结果是：一致性检查次数 $q=0.51$，运行时间 $q=0.50$，生成结点数 $q=0.48$，四者基本一致. 可以看出，4 条曲线无论从形状、相变点的位置还是变化趋势都较为吻合(除一致性检查次数在区间[0.6,0.8]偏差稍大外).

根据模型，问题无解的概率 $P_0=(1-\lambda)^r\approx e^{-S}$. 由 $P_0=0.5$ 可得交叉点的位置 $S=0.69$，由式(9)求得 $q=0.49$；相变区对应的解数区间是 $S\in[0.01,4.61]$，同样求得对应的约束强度区间是 $q\in[0.45,0.56]$. 实验的结果由表 1 给出，其中 $q\in[0.40,0.60]$区间时是 1 000 个问题的平均结果. 实验得到交叉点 $q=0.48$，相变区 $q\in[0.44,0.56]$. 两者的结果是相当接近的.

表 1　实验给出的问题集$(8,10,1,q)$上的无解比率

约束强度 q	0～42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59～100
无解千分比	0	2	20	103	177	307	516	619	770	860	909	944	969	987	991	995	998	1 000

5　指导求解的启发式

以深度优先搜索求解 CSP 时，在搜索树上应优先搜索有解并且开销小的分支. 由式(6)所示的求解开销 ζ'，可知搜索空间越小，解数越多，则分支开销越小. 在给定一个 CSP 后，问题的解数是一定的，与求解过程中变量被赋值的顺序无关，因此分支越少则每个分支的平均解数越多，搜索开销就越小. 依贪心法的思想，对给定的 CSP 和求解过程中生成的子 CSP，每次总是选取值域最小的变量进行赋值，则局部分支数最少，全局开销也可望较小. 这就对 Haralick 和 Elliot 所提出的最小值域的变量选择启发式[9]，给出了一个较好的解释.

最小值域启发式已经在文献中被证明具有较强的启发能力[11]，但也存在着一些问题：在许多情况下会有若干变量的值域大小相同，这样就只能进行随机的选取. 现在可以考虑在值域大小相同时，优先选取和其他未赋值变量约束关系最强(受约束度最大)的变量，可望在一致性实施后使其他变量的值域尽可能小，因而搜索空间也较小. 以下式计算未赋值变量 v 的受约束度，

$$t(v)=1-\prod_{u\in U-\{v\}}(1-q_{u,v}),\tag{12}$$

其中 U 是未赋值变量的集合，$q_{u,v}$是变量 u 和 v 之间的约束强度. 在随机二元 CSP 中，所有约束强度相同，所以只需考虑 U 对应的约束图中顶点的度数即可，度数大的优先选取.

对问题集$(20,10,0.5,q)$进行求解，分别使用最小值域启发式和最小值域加最强约束启发式(算法的其余部分和第 4.2 节相同)，得一致性检查的求解开销，如图 3 所示.

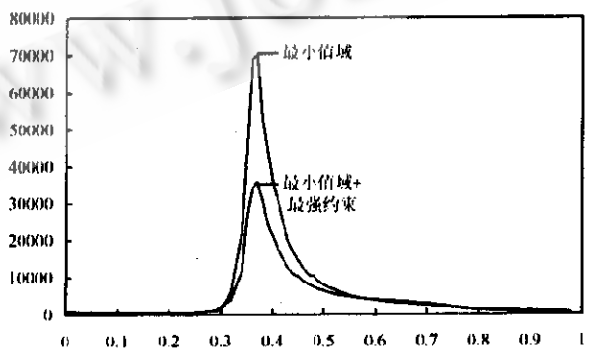

图 3

可以看出，后者的求解开销明显低于前者，尤其在相变点附近，后者的开销约为前者的 1/2. 以生成结点数衡量，有相似的结果. 但以运行时间衡量，相变点附近后者的开销约为前者的 2/3，原因一是增加了计算受约束度的开销，二

© 中国科学院软件研究所 http://www.jos.org.cn

是算法中存在着编程开销,如函数调用、输入输出等等.可以看出,启发式的贡献是压缩了开销曲线的高度,而曲线的形状和相变特性依然相似.

6 结 论

二元约束满足问题的求解开销有其内在特性,随着求解启发式的改变,开销曲线的形状不会有质的改变.因此,本文提出的以搜索结点个数评价二元CSP的求解开销的理论模型是具有一般性的,它不仅可以在相变分析中得出和实验一致的结果,而且对于解释已有的求解启发式和寻找新的更为有效的启发式提供了一定的理论工具.

参考文献

1 Detcher R, Meiri I. Experimental evaluation of preprocessing algorithms for constraint satisfaction problems. Artificial Intelligence, 1994,68(2):211～241

2 Prosser P. An empirical study of phase transitions in binary constraint satisfaction problems. Artificial Intelligence, 1996,81(1):81～109

3 Gent I P, Walsh T. Easy problems are sometimes hard. Artificial Intelligence, 1994,70(1),335～345

4 Williams C P, Hogg T. Exploiting the deep structure of constraint problems. Artificial Intelligence, 1994,70(1):73～118

5 Prosser P. Binary constraint satisfaction problems: some are hard than others. In: Cohn A ed. Proceedings of ECAI-94. New York: John Wiley and Sons Inc., 1994. 95～99

6 Crawford J M, Auton L D. Experimental results on the crossover point in satisfiability problems. In: Proceedings of the 11th National Conference on Artificial Intelligence. Menlo Park, Cambridge, London: AAAI Press/the MIT Press, 1993. 21～27

7 Smith B M, Dyer M E. Locating the phase transition in binary constraint satisfaction problems. Artificial Intelligence, 1996, 81(1):155～181

8 Gent I P, MacIntyre E, Prosser P *et al*. The constraintedness of search. In: Proceedings of the 13th National Conference on Artificial Intelligence and the 8th Innovative Applications of Artificial Intelligence Conference. Menlo Park, Cambridge, London: AAAI Press/the MIT Press, 1996. 246～252

9 Haralick R M, Elliot G L. Increasing tree search efficiency for constraint satisfaction problems. Artificial Intelligence, 1980, 14(2),263～313

10 Haselböck A. Exploiting interchangeabilities in constraint satisfaction problems. In: Bajcsy Ruzena ed. Proceedings of the 13th International Joint Conference on Artificial Intelligence. San Mateo, California: Morgan Kaufumann Publishers, Inc., 1993. 282～287

11 Frost D, Dechter R. In search of the best constraint satisfaction. In: Proceedings of the 12th National Conference on Artificial Intelligence. Menlo Park, Cambridge, London: AAAI Press/the MIT Press, 1994. 301～306

12 Frost D, Dechter R. Dead-end driven learning. In: Proceedings of the 12th National Conference on Artificial Intelligence. Menlo Park, Cambridge, London: AAAI Press/the MIT Press. 1994. 294～300

13 Smith B M. Phase transition and the mushy region in constraint satisfaction problems. In: Proceedings of ECAI-94. Amsterdam, Netherlands, 1994. 100～104

14 Sabin D, Freuder E. Contradicting conventional wisdom in constraint satisfaction. In: Alan Borning ed. Proceedings of PPCP'94: 2nd Workshop on Principles and Practice of Constraint Programming. Seattle, WA, 1994

A Node Cost Model in Solving Binary Constraint Satisfaction Problems

XUE Han-hong CAI Qing-sheng

(*Department of Computer Science and Technology University of Science and Technology of China Hefei* 230027)

Abstract A model of evaluating cost by number of nodes searched in solving binary constraint satisfaction problems is introduced in this paper. It has been applied to the random binary constraint satisfaction problems to analyze the phase transition property of solving cost. The theoretical cost induced from the model is compared with those practical costs evaluated by number of nodes searched, number of constraints checked and time consumed respectively. Based on this model, the effect of using solving heuristics to reduce cost is discussed and a new variable ordering heuristic is given.

Key words Constraint satisfaction, solving cost, phase transition, solving heuristics.

© 中国科学院软件研究所 http://www.jos.org.cn