从数据挖掘到机会/征兆发现*)

张振亚[1,2] 程红梅[3] 王煦法[4]

(中国科学技术大学多媒体计算与通信教育部—微软重点实验室 合肥 230027)[1]

(安徽建筑工业学院电子与信息工程学院 合肥 230022)[2] (安徽建筑工业学院管理工程系 合肥 230022)[3]

(中国科学技术大学计算机系 合肥 230027)[4]

摘 要 机会发现以及时发现动态确定性系统中能够对决策产生显著影响的事件为目标,是2000年以来智能信息处理研究领域中一个新兴的研究方向。本文根据机会发现的研究目标,提出了机会/征兆发现的概念。机会/征兆发现与数据挖掘和知识发现具有天然的联系,本文对机会/征兆发现与数据挖掘(知识发现)的联系与区别进行了讨论,并对文本机会/征兆发现进行了概述。

关键词 机会/征兆发现,数据挖掘,知识发现,认知

From Data Mining to Chance/Sign Discovery

ZHANG Zhen-Ya [1,2] CHENG Hong-Mei [3] WANG Xu-Fa [4]

(MOE-MS Key Laboratory of Multimedia Computing and Communication, University of Science and Technology of China(USTC), Hefei 230027)[1]

(School of Electronics and Information Engineering, Anhui Institute of Architecture & Industry(AIAI), Hefei 230022)[2]

(Management Engineering Department of AIAI, Hefei 230022)[3] (Computer Department of USTC, Hefei 230027)[4]

Abstract To discovery important event in a deterministic dynamic environment for decision-making in time, chance discovery, a new realm in intelligent information processing is focused by some researchers since 2000. According to targets of chance discovery, concept about chance/sign discovery is presented in this paper. Because there are much natural relation between data mining (knowledge discovery) and chance discovery, distinguish between data mining (knowledge discovery) and chance discovery are discussed and summarization for text chance/sign discovery is shown in this paper too.

Keywords Chance/sign discovery, Data mining, Knowledge discovery, Cognition

1 机会/征兆发现

为及时地发现动态确定性系统中对决策能够产生显著影响的事件,2000年左右,机会发现(Chance Discovery)[1~11]被提出。目前,机会发现研究认为(Ohaswa):所谓机会是指能够对决策产生显著影响的事件(event)或情形(situation)。对决策而言,一个构成机会的事件/情形或者是机遇(opportunity),或者是风险(risk)。所谓机会发现是指明晰对一个决策构成机会的事件/情形(特别是事件/情形极少发生或者意义极易被忽视时)的意义。一旦某机会被识别,若该机会意味着风险,则决策应该避免该事件/情形的出现;若该机会是成功的机遇,决策就应该促使该事件/情形的进一步活跃。对机会发现而言,机会发现不是为了预测未来要发生的事件/情形,而是要通过机会的甄别,改变或创造未来!对一个动态的系统而言,机会实际上是系统中可预见的未来系统状态变化的种子。

目前,对机会发现的定义仍然存在着争论。文[12,13]认为,机会发现的目标是确定被考察系统中事件/状态(已经发生或将要发生)的演化线索,演化线索是发生事件/状态的征兆(Sign)。在意义明确的前提下,征兆可以被认为是机会(Chance)。据此,提出了机会/征兆(Chance/Sign)、机会/征兆发现(Chance/Sign Discovery, CD)的概念。具体地[12,13]:

所谓机会/征兆(Chance/Sign),是指已经发生的事件或系统状态。作为机会/征兆的事件或系统状态能够指示已发生但尚未被意识到的事件或系统状态,或者能够预示将要或可能发生的事件或系统状态。所谓机会/征兆发现(Chance/Sign Discovery, CD),是指从大量的事件和环境信息中甄别机会/征兆的非平凡过程。

机会/征兆发现的平凡过程,是指在已经发生或来有可能发生的事件或系统状态的特征因素确定时,在被考察的事件/系统状态集中对符合确定特征条件的事件或系统状态的搜索、甄别过程。

机会/征兆发现的平凡过程,是指在关注事件/状态的特征确定时,利用这些已知的特征在数据集中搜索的过程,是一个模式匹配过程。例如,对SARS、禽流感等重大疫情的发生与流行,卫生防疫部门对一些需要关注的诸如发热等临床特征的规定以及各级卫生防疫部门对这些特征的监测。

在具体的动态确定性系统内,特定决策/任务的机会/征兆是系统内的事件,其可以指示未来将要发生(或者已经发生但尚未被意识到)的事件。通常,如果被考察系统内的事件相

*)基金项目:多媒体计算与通信教育部-微软重点实验室科研基金(05071807),安徽建筑工业学院博士后科研启动基金,安徽省教育厅自然科学项目(KJ2007A110ZC)。张振亚 博士,博士后,主要研究领域为信息检索、数据挖掘、机会/征兆发现;程红梅 讲师,主要研究领域为智能金融工程、计算机审计;王煦法 教授,博导,主要研究领域为智能信息处理、自然计算。

关的流数据(stream data)充分占有,可利用知识发现技术获取被考察系统内的知识,为积极决策提供依据。问题在于:

1) 被考察系统的事件相关流数据占有不充分时怎么办?

2) 仅仅依靠数据就可以对被考察系统进行充分刻画,近而为决策提供依据吗?

3) 对系统内事件的把握有时候依赖于人的顿悟(insight)与直觉(intuition)。

对上述问题,机会/征兆发现提供了自己的解决方案。图1给出了机会/征兆发现的基本框架[12,13]。图1中,机会/征兆的发现过程分成两部分:计算机和人。计算机对数据进行必要的分析,此时,若充分占有数据,知识发现/数据挖掘技术可使用。在计算机对数据分析完成后,分析结果供人/人群(一般为人群,该人群以发现被考察数据/系统中的机会/征兆为目标)判定机会/征兆时使用。在人/人群考察完分析结果后,或者得到结论(发现机会/征兆或者断言不存在机会/征兆),或者指示计算机进一步分析数据。一般地,需要人/人群与计算机多次交互工作才能获得结论。在人机交互过程中,对机会/征兆的认识逐步深化。显然,机会/征兆发现过程是一个人机交互作用的双螺旋过程。

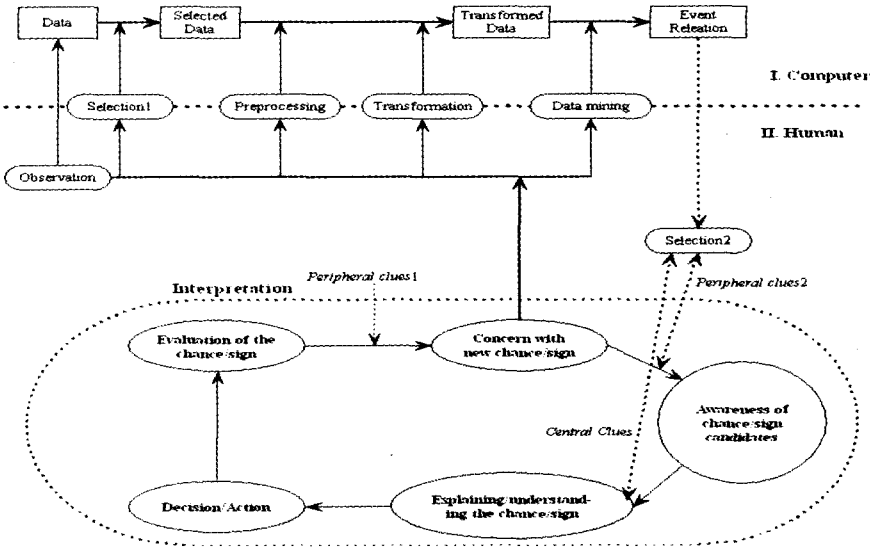

图1 机会/征兆发现的基本框架

机会/征兆的确认最终由人确定,这规定了机会/征兆发现过程中人的核心地位,同时,也为直觉(intuition)和顿悟(insight)发挥作用提供了可能。机会/征兆发现认为:机会/征兆的发现是认知活动的结果,计算机的数据分析结果只是认知活动的必要外部线索(peripheral clue)。机会/征兆发现的主要线索(central clue)是人的思维活动以及人群的社会认知。图1中,计算机提供的外部线索由peripheral clue2标识;标识为peripheral cluel的外部线索,该数据表示已知的被考察系统相关的知识/经验,这些知识/经验不是从计算机分析的数据中首先得到。

图1显示机会/征兆发现是三个部分的有机结合:第一部分(Interpretation)是针对系统内的一个具体事件/局部状态,进行机会/征兆发现的人/人群的认知活动。此时,人群之间的意见交换是不可或缺的。意见交换是人群间针对具体事件的通讯(Communication);第二部分(Peripheral clue)是外部线索。对机会/征兆发现,外部线索的注入,可以为机会/征兆发现提供必要的场景信息,同时,也给人/人群注意的迁移提供了外部环境;第三部分是针对数据分析任务的数据建模与挖掘(Data modeling&Mining)。对特定的领域,需提供适合机会/征兆发现需要的新的数据建模和其上的数据分析手段。

机会/征兆发现以发现数据集中蕴涵的可能成为机会/征兆的事件/状态为目标,以期为积极决策提供必要的支持,这与数据挖掘、知识发现有着类似的目标。同时,由于都需要对被考察系统内的流数据进行分析,机会/征兆发现与数据挖掘、知识发现有着天然的联系。本文的第2、3部分对机会/征兆发现与数据挖掘和知识的联系与区别进行了讨论,第4部分概述了文本机会/征兆发现。研究展望在最后给出。

2 机会/征兆发现与数据挖掘

数据挖掘的中心任务是对海量的数据进行处理。数据挖掘在处理数据时,在目标明确(挖掘确定范畴内的知识)的前提下,选用合适的数据挖掘算法,对数据的固定特征进行分析。以C4.5算法为例:其对数据的固定的属性集合进行考虑,而对该固定属性集合以外的数据的属性不考虑。

设$Y=f(X)=f(x_1\cdots x_n)$规定了一个动态确定性系统,即对任何一个对系统f的输入$X=(x_1\cdots x_n)$,x_i是f的一个特征,f可以产生唯一的一个响应[2]。在这样一个系统中进行数据挖掘的目标是通过该系统中海量的数据分析获得系统f的一个客观描述。除非全面占有数据,否则数据挖掘只能对f进行基于部分特征的近似描述!对f,机会/征兆发现与数据挖掘不同,机会/征兆发现的目标不是获得系统f的描述,而是为了获得系统f已经获得的输入以及系统f的某个局部状态,这些输入/局部状态可能对系统f的未来状态产生重要影响,或者f中其它未被意识到的局部状态因为这些输入/局部状态已经显著地改变。机会/征兆发现在对f中状态或事件的分析过程中,系统的大量特征信息可以是未知的。

机会/征兆发现很容易和预测混淆。以f为例,预测的目标是为了对动态系统f进行确定的认识,并根据认识对f中的未来状态或者未来某个局部状态进行断言。而机会/征兆强调在对f未知或不完全清楚的情况下,对f中的某个事件的意义进行断言。

可以从海王星发现过程[14]辨析机会/征兆及机会/征兆发现与预测的差异。1781 年，天王星被发现后，人们注意到天王星的运动轨道总是偏离天体力学计算的轨道，于是便推测天王星轨道之外可能还存在一颗行星，它的引力作用使天王星的轨道运动受到“摄动”。1845 年到 1846 年，英国的亚当斯和法国的勒维耶这两位年轻人，根据牛顿万有引力和运动定律，分别独立进行了计算，他们反过来从天王星运动的偏差去估计摄动的大小，从而推算出未知行星的位置。依照勒维耶的计算结果，1846 年 9 月，柏林天文台的天文学家果然在预测位置附近发现了海王星！从海王星的发现过程可以看出，天王星的发现以及天王星运行轨道的异常是海王星被发现的征兆；而利用天王星运行轨道异常的数据以及牛顿定律对海王星出现的位置的确定则是具体的预测。此时，预测分析的任务的核心是对牛顿定理对天体系统实用性的认识，而机会/征兆发现只是找到天王星以及天王星运行轨道异常这一现象。

3 机会/征兆发现与知识发现

Fayyad 将知识发现(KDD)[15]定义“从数据集中识别出有效地、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程”。知识发现的框架流程可由图 2 示意。

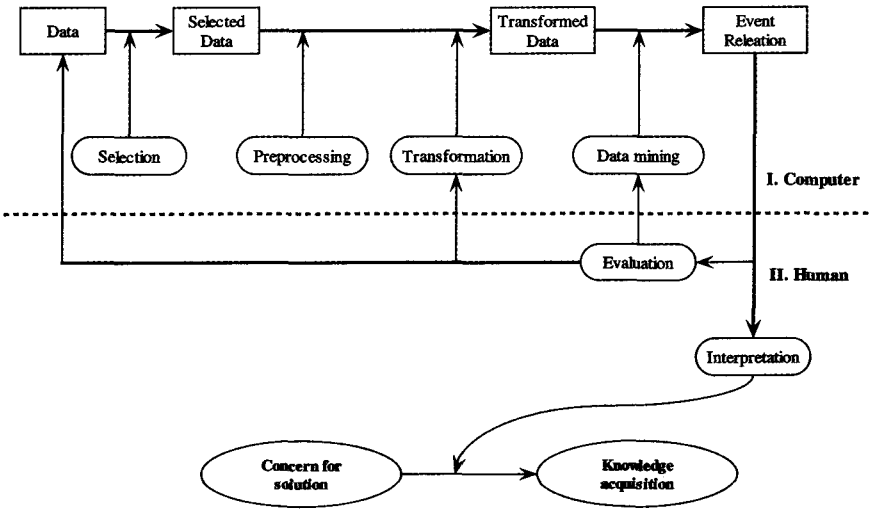

图 2 KDD 的框架流程

图 2 给出了知识发现的框架流程。图 2 显示，为发现知识，KDD 可以通过评估(Evaluation)机制对数据分析过程进行一定程度的干涉。由于 KDD 的目标是通过被考察系统内的海量数据进行自动/计算机辅助分析获得系统某些可描述特征的描述，评估机制只是 KDD 完成任务过程中的一种外部线索(Peripheral clue)。对比图 1 示意的机会/征兆发现的基本框架，机会/征兆发现将数据分析/计算机辅助分析的结果作为一种外部线索，机会/征兆发现是以外部线索为起点的认知活动：认知活动可能围绕被分析的数据，但更多的可能是被分析的数据仅仅是一个起点！这是机会/征兆发现与 KDD 最大的差异。如果将以知识发现中数据分析结果为外部线索的机会/征兆发现过程作为知识发现对数据分析结果的评估过程的实现，由于机会/征兆发现过程中认知心理活动的参与，知识发现的评估过程可以更加拟人化进行。此时，机会/征兆发现是 KDD 的必要的、有益的补充[13]。

4 文本机会/征兆发现

随着信息技术的飞速发展，Internet 已经成为传统的报纸、广播、电视之外的第四种重要的信息发布形式。与其它三种媒体发布信息的形式相比，Internet 上信息的海量、开放等特性给信息的有效获取提出了新的挑战。Internet 上的海量信息中，文本是一种重要的信息承载形式，其中蕴涵着众多有价值的信息。Internet 上传播的文本信息获得的方便性为文本信息处理研究提供了极大的便利。从文本信息中有效地获取高价值的信息是文本机会/征兆发现研究关注的内容。目前，以文本信息为研究对象的机会/征兆发现研究主要集中于关键字获取、主题发现与传播等方面。机会/征兆发现技术应用于文本高价值信息的获取可以有效地把人对文本信息内容的认知融入文本信息的分析过程。

目前，利用机会/征兆发现技术获取文本信息中的关键字研究的典型方法是 KeyGraph[2,4] 和 KeyWorld[2,16]。KeyGraph 在词频统计的基础上，通过对候选关键字进行关联性分析、分析过程可视化以及人机交互等技术的运用，较好地实现了关键字的抽取。KeyWorld 对学术论文形式的文本信息进行了以小世界模型为基础的可视化组织，实现了以小世界模型的特征路径长度以及聚类系数为基础的关键字自动抽取。KeyGraph 与 KeyWorld 在实现文本信息关键字抽取都以文本信息的高频词为基础，通过各自的选择标准，从非高频词中选取部分词与高频词共同构成以候选关键字为节点的图；对候选关键字，两种方法依据节点对全图的贡献的大小选择一定数量的候选关键字词作为关键字。与词频统计分析相比，两种方法获取的关键字较好了表达了文本信息的内容。机会/征兆发现研究中，KeyGraph 还被有效地用于 WWW 中的主题发现与传播[7]。

综合目前机会/征兆发现研究中的对文数据的分析处理，基本思路是：1)首先根据任务确定信息的基本单元；2)其次以基本单元为节点，结合特定的领域知识构造相关的描述图；3)进一步，结合用户的认知，描述图不断进化；4)最后对通过描述图的状态有效的解析获取确定的结论。上述流程中，步骤 2)、3)是机会/征兆发现对文本信息处理的关键[12,13]。

人对文本信息的理解，一方面依据文本信息的内容，另外一个方面依赖于人对文本信息内容的认知。目前对文本机

会/征兆发现的研究，用户的认知主要通过对文本信息内容的计算机辅助分析过程中通过交互实现。这种认知的作用方式，忽视了注意作为人意识心理活动信息过滤器[17~31]应该发挥的作用，这是信息科学领域机会/征兆发现研究的严重不足。同时，被考察系统的时间属性被忽略是目前进行的机会/征兆发现研究的一个特点，也是机会/征兆发现研究的另一个不足。

研究展望 智能信息处理领域中，机会/征兆发现是一个新兴的研究方向，目前进行的研究仅仅处于起步阶段。由于对机会/征兆的发现是心理活动的结果，认知心理和社会心理是机会/征兆发现的机制、模型构造时必须考虑的因素，同时，相关的方法研究也不能够忽视认知心理和社会心理的作用。

目前，在报纸、广播、电视、Internet 等重要的媒体形式中，文本是信息的重要载体。利用基于认知的文本机会/征兆发现方法，及时有效地发现蕴涵于文本信息中的机会/征兆，对政府部门而言，可以为对关系国计民生的决策提供依据；对商业机构和生产商而言，可以指导其商业行为；对个人而言，可以助其把握机遇，避免风险。同时，在信息科学领域，机会/征兆发现是一个新兴的方向，基于认知心理的文本机会/征兆发现方法研究所取得的进展将促进机会/征兆发现研究的成长，具有深远的理论意义和重大的现实意义。

参 考 文 献

1 Zimmerman C. The development of scientific reasoning, Developmental Review, 2000,20:99～149
2 Yukio O, Peter M, eds. Chance Discovery, Springer- Verlag, 2003
3 Ohsawa Y, Nara Y. Understanding Internet Users on Double Helical Model of Chance-Discovery Process. In: Proc. of IEEE. International Symposium on Intelligent Control, 2002. 844～849
4 Ohsawa Y, Benson NE, Yachida M. KeyGraph: Automatic Indexing by Cooccurrence Graph Based on Building Construction Metaphor. In: Proceedings of Advances in Digitial Libraries Conference (IEEE ADL's 98). 12～18
5 Ohsawa Y, Yachida M. Discovery Risky Active Faults by Indexing an Earthquake Sequence. In: Proc. International Conference on Discovery Science, 1999. 208～219
6 Ohsawa Y. Chance Discovery for Making Decisions in Complex Real World. New Generaion Computing, 2002, 20(2):143～163
7 Takama Y, Hirota K. Discovery of Topic Distribution through WWW Information Retrieval Process. In: Proc. of IEEE(2000). pp1644～1647
8 Goldberg DE. The Design of Innovation: Lessons from and for Competent Genetic Algorithms, Kluwer Academic Publishers, Boston, MA, 2002
9 McBurney P, Parsons S. Chance Discovery Using Dialectical Argumentation. In: JSAI 2001 Workshops, LNAI, 2253. 414～424
10 诸世卓，陈小平. Agent 机会发现的一种相关性描述. 计算机工程与应用，2004,5:45～48
11 诸世卓，陈小平，皮亮. Agent 机会发现的一种刻画：溯因推理及其扩展. 计算机工程，2004,30(12):40～42
12 高俊波. 基于认知的征兆发现理论和方法研究：[博士学位论文]. 中国科学技术大学档案馆，2005
13 张振亚. 从数据挖掘到机会/征兆发现：[博士后研究报告]. 中国科学技术大学联想实验室资料室，2006
14 地球科学. http://www.nju.edu.cn/njuc/dikexi/earthscience/chp1/3-2-9.htm
15 Fayyad U M, Piatetsky-Shapiro G, Smyth. From data mining to knowledge discovery. In: Fayyad. U. M Piatetsky-Shapiro G., Smyth P, Uthurusamy R, eds. Advances in knowledge discovery, data mining, AAAI Press/MIT Press, CA, 1996. 1～31
16 Yutaka M, Yukio O, Mitsuru I. KeyWorld. In: Extracting Keywords from a Document as a Small World, Proceedings the Fourth International Conference on Discovery Science, 2001. 271～281
17 孟昭兰主编. 普通心理学. 北京大学出版社，2003
18 Zimbardo P. Psychology and Life. Illionis: Scottand Forseman and Company, 198519) Catherine M A, Thomas H C, Andrew R M, et al. Neural mechanisms of visual attention: Object-based selection of a region in space. Journal of Cognitive Neuroscience, 2000, 12(Suppl. 2):106～117
20 Treisman A, Gelade G. A feature-integration theory of attention. Cognitive Psychology, 1980,12(1):97～136
21 王健，朱祖祥. 视觉注意选择性的认知心理学理论研究进展. 应用心理学，1997,3(1):58～64
22 Posner M I, Synder C R, Davidson B J. Attention and the detection of signals. Journal of Experimental Psychology: General, 1980,109(2):160～174
23 Tsal Y. Movement of attention across the visual field. Journal of Experimental Psychology: Human Perception and Performance, 1983,9(4):523～530
24 杨华海，赵晨，张侃. 外源性视觉选择性注意的时空特征。心理学报，1998,30(2):136～141
25 Eriksen C W, Murphy T. Movement of attentional focus across the visual field: Acritical look at the evidence. Perception and Psychophysics, 1987,42(3):299～305
26 Larberge D, Brown V. Theory of attentional operation in shape identification. Psychological Review, 1989,96(2):101～124
27 Egly R, Driver J, Rafal R D. Shifting visual attention between objects and location: Evidence from normal and parietal lesion subjects. Journal of Experimental Psychology: General, 1994,123(2):161～177
28 傅世敏，陈霖. 对"物体内注意转移"优势效应之机制的进一步检验. 心理学报，1999,31(2):142～147
29 Driver J, Baylis G C. Movement and visual attention: The spotlight metaphor breaks down. Journal of Experimental Psychology: Human Perception and Performance, 1998,15(3):448～456
30 Tsal Y, Lavie N. Location dominance in attending to color and shape. Journal of Experimental Psychology: Human Perception and Performance, 1993,19(1):131～139
31 刘志华，陈彩琦，金志成. 选择性注意的理论及其发展趋势—认知神经研究. 心理科学，2003,26(4):709～712

(上接第 176 页)

目前，我们仍在深入研究 NT 算法和皇冠分解在实际中的应用，期望能够通过二者的有机结合改进特定情况下参数化点覆盖问题的时间复杂度下界。

参 考 文 献

1 Garey M, Johnson D. Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco, 1979
2 Downey R G, Fellows M R. Parameterized Complexity. New York: Springer, 1999
3 Buss J F, Goldsmith J. Nondeterminism within P. SIAM Journal on Computing, 1993, 22 (4): 560～572
4 Downey R G, Fellows M R. Parameterized computational feasibility. In: Clote P, Remmel J, eds. Feaible mathematics II, Boston, Birkhauser, 1995. 219～244
5 Balasubramanian R, Fellows M R, Raman V. An improved fixed parameter algorithm for vertex cover. Information Processing Letters, 1998, 65(3): 163～168
6 Stege U, Fellows M. An improved fixed-parameter- tractable algorithm for vertex cover: [Technical Report]. Department of Computer Science, ETH Zurich, 318, 1999
7 Niedermeier R, Rossmanith P. Upper bounds for vertex cover further improved. Lecture Notes in Computer Science, 1999, 1563: 561～570
8 Chen J, Kanj I A, Jia W. Vertex cover: Further observations and further improvements. Journal of Algorithms, 2001, 41(2): 280～301
9 Abu-Khzam F N, Collins R L, Fellows M R, et al. Kernelization algorithms for the vertex cover problem: theory and experiments. In: Proceedings, Workshop on Algorithm Engineering and Experiments (ALENEX), 2004
10 Fellows M. Blow-Ups, Win/Win's, and Crown Rules: Some New Directions in FPT. Lecture Notes in LNCS, 2003, 2880: 1～12,
11 Chor B, Fellows M, Juedes D. An efficient FPT algorithm for saving k colors. Manuscript, 2003. 7
12 Chlebik M, Chlebikova J. Crown reductions for the Minimun Weighted Vertex Cover problem. Electronic Colloquium on Computational Complexity, Report No, 2004
13 Nemhauser G L, Trotter L E. Vertex Packings: Structural properties and algorithms: Math Programming, 1975, 8: 232～248
14 Chen J. Parameterized Computation and Complexity: A New Approach Dealing with NP-Hardness. Journal of Computer Science and Technology, 2005, 20: 18～37