Web使用挖掘系统研制中的主要问题和应对策略

张 锋 常会友
(中山大学信息科技学院计算机系 广州510275)

Key Issues and Solution Strategy in R&D of Web Usage Mining Tools

ZHANG Feng GHANG Hui-You
(Department of Computer, College of Information Science and Technology, Zhongshan University, Guangzhou 510275)

Abstract With the rapid development of WWW, Web Usage Mining, as well as Web Mining, has become a hot direction in academic and industrial circles. It is generally believed that there are three tasks, preprocessing, knowledge discovery and pattern analysis, in Web Usage Mining. Though Web Usage Mining is still ranged in the application of traditional data mining techniques, in view of changes in application environment and operated data concerned, some new difficulties have arisen accordingly. This paper takes efforts to address such challenges in the three phases and introduces some proposed solutions simultaneously.

Keywords Data mining, Web mining, Web usage mining

1. 引言

Internet的蓬勃发展促使Web数据挖掘在学术界和工业界成为一个热点。Web站点中主要包括有三大类数据:网页内容数据、Web结构数据和记录访问者浏览网站的日志数据,相应地对这三种数据进行数据挖掘操作,Web数据挖掘也就分为三类:Web内容挖掘、Web结构挖掘和Web使用挖掘[1]。

Web内容挖掘和Web使用挖掘是两类相关联的Web数据挖掘研究。文[2]介绍了Web内容挖掘的基本概念。文[3,4]则提供了一些跟Web结构挖掘有关的有用信息。但这两方面的内容都超出了本文讨论的范围。下面我们集中讨论Web使用挖掘。

Web Server每次收到一个用户请求的时候网站都会自动在日志文件记下一个事务项。主要有三类的日志文件:访问日志,引用日志和代理日志。如果是按这么分的话,这里的访问日志文件是CLF(Common Log Format)格式的,格式结构如下:

HostID rfc931 authuser[date offset] "method URL protocol" status bytes

另一种格式的日志文件叫ECLF(Extended Common Log Format)。就是在CLF后添加两个数据域Referrer(导航到当前页的URL)和Agent(用户的操作系统和浏览器),域含义如表1所示[30]。如果没有特别说明,我们在论文的以下部分提到的访问日志文件都是ECLF格式的。

Web使用挖掘用到的数据包括服务器日志文件、用户档案数据、注册数据、Web内容和Web结构信息等等。

Web使用挖掘主要是利用这些数据分析用户的浏览行为或者说浏览模式。Web使用挖掘得出的结果主要有两方面的作用:一是帮助网站作出正确高效的市场营销策略;二是优化整个网站的逻辑结构,以给用户提供更加有效的个性化服务和高效的网站—用户沟通渠道[7,8]。

表1 ECLF域描述

域	含义
HostId	远程机器名或IP地址
Rfc931	用户登录名称
AuthUser	服务器授权用户名
Date	请求日期和时间
Offset	当地时间与格林威治时间偏移值
Method	请求方法(Get, Post, Head等)
URL	请求页面完整地址
Protocol	客户端使用的http通讯协议
Status	http服务器处理结果状态
Bytes	传输字节数
Referrer	指向当前请求的链接URL
Agent	客户使用的操作系统和浏览器

虽然Web使用挖掘系统在具体实现时采用的结构和技术千差万别,但其要实现的功能还是可以大致分为三部分,预处理、知识发现,模式分析,如图1所示[11]。文[7,8]中描述的WebMiner正是这样的一种典型结构的使用挖掘系统。

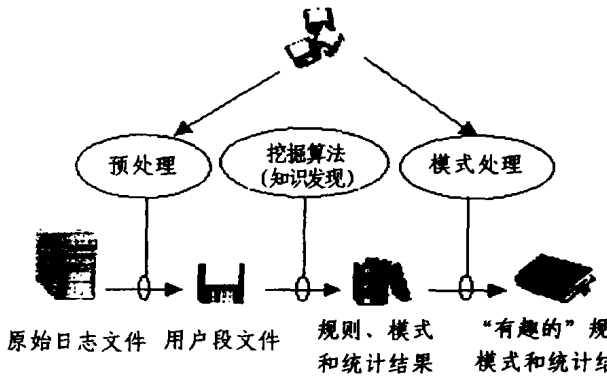

图1 Web使用挖掘总体结构图

在这三个阶段中,都分别会有一些需要解决的问题,这些

张 锋 硕士研究生,研究方向为数据库,数据挖掘,WWW数据挖掘。常会友 教授,博士,主要研究领域为数据库,网络,企业信息管理和改造,MRP和ERP。

问题很多是Web使用挖掘工具研发的瓶颈。下面我们讨论这些主要问题并总结一些已经存在的解决方案。

2. 预处理

预处理要用到的数据除了日志文件，还可以包括网页、网页结构、用户档案信息和登录信息等。Web使用挖掘的任务包括数据清理、用户识别、段识别、路径修补和事务识别等[9]。其它像概念层次的建立和URL分类在某些情况下也会用到[6]。这一阶段的目的是为知识发现阶段提供结构化、可靠和完整的数据源。

它首先要做的一件事就是进行数据清理。一些无关的日志项对分析用户的浏览行为会产生误导，数据清理就是把日志文件中一些与数据分析和挖掘无关的项清除。首先要做的就是把日志文件中标识失败(failor，error)的项清除；二是要把一些自动搜索代理(比如象很多搜索引擎的"网络蜘蛛"Crawler或Spider)产生的日志项辨别出来并删除，因为这些数据并不是由访问用户产生，对Web使用挖掘技术分析用户浏览行为是没有帮助的。许多网络蜘蛛访问网页的时候都会自动地在访问日志的agent域里做登记，这样就很容易根据这个信息把这些日志项辨别并删除。另外，为了剔除这些日志项，也可以用辨别"非人为的请求日志项"的方法，文[6]提出了三个辨别标准：1)来自同一个主机的对同一个URL的访问，2)请求之间的时间间隔太短，并不足够访问页面内容，3)来自于一个主机的访问URL的引用URL是空值，因为引用URL为空只能说这个访问是直接敲入URL，或者是用书签访问，又或者是用脚本来访问的；三也是一个比较有争议的问题—是否应该把包含图像音频视频文件的日志项清除或者怎样删除。因为在这些日志项里面，有一些可能确实是用户浏览的，有一些却是用户浏览页面的时候系统自动下载的。在WebMiner中[7~9]是把它清除掉的，而在原型系统WebLog-Miner[14]中，是把这些数据保留的，但需要指出的是它仅仅是一个Web日志OLAP原型系统。一般在浏览模式挖掘的系统中，是把图像项删掉的，因为即使这些日志项确实是用户访问产生的，它们对用户的访问行为分析影响也很小。

下一个工作是用户识别，不同用户使用一台主机或者是不同用户使用一台代理上网的时候，都会存在着用户识别的问题，因为日志文件只是记录主机或代理服务器的IP地址。简单的解决方法是使用Cookie或者是授权机制，文[16]中也提出了一个叫Remote agent的跟Cookie类似的机制，但是这些方法都需要用户的合作，而往往用户会觉得麻烦或者从隐私角度考虑，不会给予配合。

文[9,15]提出了一些启发方法。比如它认为使用不同的浏览器或不同操作系统的用户(这些信息记录在日志文件的agent项)访问来自不同的用户。这是有一定合理性的，因为一个用户很少会同时使用不同的浏览器来访问网站，更别说同时使用不同的操作系统了，但如果用户确实这么做的话，这种区分方法就会产生误导；另外它认为来自同一个IP的一个页面访问没能够在已访问的页面找到一个指向它的连接的时候，就是一个新的访问用户，比如一个访问序列(A，B，C，D)，如果A有指向B的连接，B有指向C的连接，但ABC都没有指向D的连接，就认为ABC是来自一个用户，而D是另一个用户的页面请求，但如果用户是直接敲入URL或者用书签来访问的话，这个方法也无能为力了。

做完用户识别，下一个是用户段识别。所谓用户段就是同一用户在一段时间内访问的连续的网页集。跟用户识别类似，用Cookie或Session机制也可以实现段识别，但是这也存在着与用户识别一样的问题。现在普遍使用的方法是：a)看整个页面集的访问总时间是否超过了限定值；或b)两个连续的访问页面的时间间隔是否超过限定值。一般认为b)更加符合实际一点。WUM和WebMiner都使用b)方法来标识段，这个时间一般认为是30秒或25.5秒[13]。

接下来要做的是路径修补。它要确认用户段的网页访问序列中是否有重要的页面访问是没有登记的，这个问题产生的原因是Cache的存在。当用户在访问网页过程中用了后退按钮的时候，前面的页可能有一个拷贝在Cache中了，于是客户机从cache直接调出这个页面拷贝，也就导致这个页面访问不会登记在访问日志中了。路径修补正是要把这些不完整的路径修补完整。这个工作需要用到引用日志(或ECFL的引用域)和网络拓扑图。如果一个页面的引用URL并不是它的直接前面页面的话，说明路径"不完整"，于是就检查这个引用URL在当前用户段中是否已被访问过。如果没被访问过的话，正如在用户标识中指出那样，是一个新的用户标识；如果引用过的话，说明用户用了回退健操作，可以根据引用日志和网络拓扑结构图提供的启发信息把路径修补完整[9]。

预处理的最后一步应该做事务识别了，事务识别是和要"挖掘"出什么样的知识有关。要挖掘出什么知识，就相应地要作出跟这种知识相对应的事务定义。比如在WUM[6]中，用户段是等同于事务定义的。它把我们上面提到的用户段定义认为是"基于持续时间"的用户段定义，而这里的事务则认为是"基于结构"的用户段定义。由于WUM的目的是分析用户的浏览模式，因此前一种方法它认为是适合的；而由于WebMiner[9]是对访问日志做关联规则和序列分析，因此它还会在用户段的基础上分出更小粒度的"事务"，就是语义相关的访问页面集合。

为了把用户段分为事务，WebMiner把用户段内的网页分成了辅助页和内容页。所谓辅助页就是为了方便用户浏览信息的页面，它在事务中的作用主要是提供链接导航用户访问内容页面。内容页用户最终想访问的页面，包含一系列的信息。基于辅助页和内容页的定义，事务的定义也就有了两种：一种是所有的辅助页和它们最终到达的内容页组成一个事务，叫辅助—内容事务。对这类事务的挖掘分析可以得出一般的对一个内容页的访问模式。二是把所有的内容页面看成一个事务，对内容事务的挖掘会得出内容页面的关联信息。基于这个定义，WebMiner提出了两种识别事务的方法，一是reference length，二是maximal forward reference，两种方法识别出语义相关的事务，而且提出一种叫Time window的机制作为比较前两种方法的基准[9]。

3. 知识发现

在知识发现阶段要解决的问题是找出什么样的知识，在这一节里我们主要看看Web使用挖掘可以得出什么样的知识。早期的日志分析工具，只是提供一个报告用户活动的有限机制，实际上也就是做了一些统计工作，比如页面点击数和访问时间，最常访问的URL，每月/年的点击数等等。严格来说这只是一些日志文件的分析工具，并不能称为挖掘工具。

文[18]提出一种命名为"PageGather"的基于聚集的算法，它其实就是依据内容把用户的访问页分类，目的是根据不同的用户群体划分出它们感兴趣的网页集。但它并不关心这

些页面用户是怎么浏览页面的。

其实文[18]和文[26～29]一样，都是 Web 使用挖掘方面被称为 Personalization 的一大类应用。虽然他们实现机理不尽相同，但他们的共同之处都是基于访问日志基础上把用户按照不同的访问行为进行聚集建模，通过这个模型给新用户指导信息，帮助新用户浏览网站。这已经有点 Web 内容挖掘中的功能了。

WebLogMiner[5]根据日志文件构造一个数据立方体，再对数据立方体做传统的象 Roll-up 和 Drill-down 的 OLAP 操作或者传统的数据挖掘的操作，比如数据特征描述、类比较、关联规则、预测、分类、时间序列分析等，实际上是一个对 Web 服务器日志文件应用数据挖掘技术的汇集。而 WebMiner[7～9]则是经过了一系列的预处理工作后，把访问日志整理为一个个语义相关的事务，然后做关联规则和序列模式的数据挖掘，它被认为是传统数据挖掘技术在 WWW 使用挖掘中的应用。

MiDAS[12]和 WUM[6,13]都认为把浏览模式纯粹地看作关联规则或传统序列把问题过于简单化了，而且把访问次数作为浏览模式挖掘的一个衡量标准也不是很准确的。

所以在 MiDAS 中把传统的序列模式做了一个扩展定义，这个扩展定义被称为浏览模式，而且为了找出更有效的序列模式，被它称为“领域知识”的一些信息被引入 Web 挖掘中。四类与 Web 有关的领域知识被引入，它们是导航模板、网络拓扑结构，概念层次和句法约束。

WUM 还是认为 MiDAS 中的关于浏览模式的定义是传统的，因为它的本质还是访问一系列与时间相关的网页 URL。WUM 不再认为访问模式是一维矢量结构的模式，而把它定义为一个有向非循环图。它其实是把预处理得到的事务构建一个 Aggregate tree 的结构作为挖掘的对象，挖掘浏览模式的时候，通过一个叫 MINT 的数据挖掘语言，施加一些约束条件，从 Aggregate tree 中找出让用户感兴趣的浏览序列。

4. 模式分析

模式分析的主要任务是过滤掉没有用的信息并对感兴趣的模式进行可视化和解释的工作。

现在许多的 Web 使用挖掘原型系统象 WUM、WebMiner、MiDAS 都嵌入了语法跟 SQL 有点类似的数据挖掘语言，这样的挖掘语言首先就可以提供一些像支持度、可信度这样的客观标准，过滤掉一些被认为是不重要的知识。另外像用 OLAP 对数据立方体的分析也有助于从多个角度去理解得到的数据。可视化技术也可以帮助分析员更好地理解数据模式和预测数据的走向。内容和结构信息同样有助于过滤掉一部分无用信息。

VibViz[19]在 Web 访问模式的可视化方面做了一些前瞻性的工作。它利用 Web path paradigm 来提取用户访问模式。而且 WebViz 把用户的浏览模式定义为一个有向循环图，节点是访问的文档(页面)，节点之间的链接表示页面之间的超链。

在模式分析中另一个重要的问题是找出所谓“有趣的”模式，其实这也是传统知识发现领域一个重要的课题。很多文献涉及到这个问题。

判断挖掘出的知识是否“有趣”的标准一般可以分为两大类：一是用所谓的“客观”标准，就是根据规则结构和所用的数据确定的标准，像可信度和支持度之类就属于“客观标准”。二是所谓的“主观标准”，不仅基于规则结构和所用数据，而且和分析模式的用户有关，这些标准可能对一些用户是有趣的，但对另外一些用户却不是有趣的。

可以通过和已存在的规则集(一般叫信任集 belief)进行比较来确定一个新得到的规则是否有趣。

文[20,31]提出了一个独立于具体应用的判断模式是否有趣的两个标准：Actionable 和 Unexpected。Actionable 标准是说新发现的规则是可以为用户所利用的；而 Unexpected 标准是说新规则是用户预先不知道的或者说是与用户已知的知识冲突的。一个模式是否有趣是根据它怎么影响信任系统或与信任系统抵触来衡量。一个与 Hard 信任系统抵触或对 soft 信任系统的影响超过一定阈值的模式被认为是有趣的。而且对一个新规则 p 的“有趣度”，它用以下的这样一条公式来表示：

$$I(p,B)=\sum_{\alpha\in B}\frac{|P(\alpha|p,\zeta)-P(\alpha|\zeta)}{P(\alpha|\zeta)}$$

其中 B 是信任集系统，ζ 是支持 α 的已经发现的证据，而且它是这样定义 $P(\alpha|p,\zeta)$ 的：

$$P(\alpha|p,\zeta)=\frac{P(\alpha|\alpha,\zeta)P(\alpha|\zeta)}{P(p|\alpha,\zeta)P(\alpha|\zeta)+P(p|\rightarrow\alpha,\zeta)P(\rightarrow\alpha|\zeta)}$$

文[20]建议除了 bayesian，还可用像 dempster-shafer，frequency，cyc 和 statistic 等方法来衡量信任集的可靠强度。

文[21]利用了模糊匹配技术来比较新产生的规则和已存在的规则集。已存在的规则集用模糊集理论来描述，而新的规则则用模糊匹配技术来与已存在的模糊规则集比较。实际上在这篇论文中，它使用一个规则距离(distance)的概念来描述新产生的规则是否是 unexpected 的，也就是是否“有趣的”。它认为如果新产生的规则和已存在的信任集前提条件相同但结果不同(被称为意外的结果)，或者是前提条件不同但结果相同(意外的条件)，则认为新规则和已存在的结果是不同的，也就是有趣的。但与文[20,31]相比这种方法的缺陷在于：一些没有与信任集的条件或结果冲突的新规则，这种方法并不能发现但实际上它们是有趣的。

文[22]使用一种叫“General impressions”的方法来描述已存在的规则集(信任集)特征，GI 实际上是用户对领域知识的认识不是很精确但有模糊认识时的一种表达方法。文中给出了 GI 的两种定义：Type_1 GI 和 Type_2 GI。而且分别针对这两种定义给出了发现和评估“有趣”规则的算法。它认为“证明”当前规则集的规则不是有趣的，而与 GI 相抵触的规则看成是“有趣”的，另外，不包含在 GI 的即是 GI 中没有考虑到的规则也是“有趣”的。

文[35]则进一步加强了文[22]中的工作，它实现了一个叫 IAS(Interestingness Analysis System)的系统。它对领域知识的描述使用的三种规范。除了文[22]中的 GI 外，还包括 reasonably precise concept 和 precise knowledge。并且针对这三种知识描述规范找出“意外的”规则的方法。IAS 还包括一个对得到的结果规则集可视化的子系统。

以上的方法基本上都是从句法上比较新规则和已存在规则以确定新规则是否“有趣”。而[23,33]从逻辑上比较新规则和已存在规则集是否矛盾来确定新规则是否“有趣”，并且给出了找出 Unexpected 规则的算法。而且它定义一个规则 A→B 是与信任规则 X→Y 比较是 Unexpected 的，那么以下条件成立：

(a)B AND Y |=FALSE,也就是说 B 和 Y 逻辑上是矛盾的。

(b)A AND X 在数据集 D 的一个统计上足够大的元组子集上成立。

(c)A,X→B 成立也就是说 A,X→～Y 成立。

文[34]则扩展了文[23,33]的工作,引入了最小意外模式集的概念,并详细给出了实现的算法。找出了更符合要求更精简的规则集。

文[24,25]构造了一个叫 WebSIFT 的系统,其实它的基本结构是建立在 WebMiner 上的。但它跟 WebMiner 最大的改进在于它基于支持集理论建立了一个判断什么样的知识是"有趣的"的机制,并且实现了这样的算法。

文[22,23,33]虽然给出了手工建立的"信任集"的例子,但都没有实现从大量的数据中自动建立现实的"信任集"的算法。WebSIFT 利用网页内容和结构信息,通过支持集理论自动建立"信任集",并实现了基于信任集选出有趣的知识的算法。

结束语 我们花了很大工夫来指出 Web 使用挖掘中面对的难题,并总结了一些解决方案。其实 Web 使用挖掘作为一个新的研究方向,很多问题都在探索阶段,是没有结论的,很多学者都提出自己的解决方案。比如说在数据预处理需要的准确的访问日志数据和用户的隐私考虑之间的矛盾使得用 Cookie 等的机制来辨识用户的方法不是普遍适用的。所以得考虑其它的方法,但没有一种方法是可以完美地解决存在的问题,究竟使用哪种方法,就得综合考虑应用环境、费用等等因素。而在知识发现阶段,我们主要关注可以得到什么样的知识,这是有效地挖掘出知识的问题,但还有一个高效地挖掘知识的问题,这就涉及到挖掘算法和挖掘系统的结构设计等诸多问题。随着技术的不断向前发展,这方面的研究也会跟着不断深入。

参 考 文 献

1 Madria S K, Bhowmick S S, Ng W K, Lim E-P. Research issues in web data mining. In: Proc. of Data Warehousing and Knowledge Discovery, First International Conference, DaWaK'99. 1999. 303～312

2 Huang Lan. A Survey On Web Information Retrieval Technologies. http://citeseer.nj.nec.com/336617.html

3 Chakrabarti S, et al. Mining the web's link structure. COMPUTER, 1999, 32: 60～67

4 Kleinberg J M. Authoritative sources in a hyperlinked environment. Journal of ACM, 1999, 46: 604～632

5 Zaiane O R, Xin M, Han J. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In: Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, Ca, Apr. 1998. 19～29

6 Berendt B, spiliopoulou M. Analyzing navigation behavior in Web sites integrating multiple information systems. VLDB Journal, Special Issue on Databases and the Web, 2000, 9(1): 56～75

7 Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the world wide web. In: 9th. Conf. on tools with AI, Dec. 1997

8 Mobasher B, Jain N, Han E, Srivastava J. WEBMINER: Pattern Discovery from World Wide Web Transactions: [Technical Report TR96-050]. Department of Computer Science, University of Minnesota, Feb. 1996

9 Cooley R. Mobasher B, and Srivastava J. Data Preparation for mining world wide web browsing patterns. Journal of Knowledge and information System, 1999, 1(1): 5～32

10 Pitkow J. In search of reliable usage data on the WWW. In: Sixth Intl. World Wide Web Conference, Santa Clara, CA, 1997. 451～463

11 Srivastava J, Cooley R, Deshpande M, Tan P N. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD explorations, 2000, 1: 12～23

12 Buchner A G, Baumgarten M, Anand S S, Mulvenna M D, Hughes J G. Navigation pattern discovery from internet data. In: Proc. of the 5th ACM Intl. Confe. on Knowledge Discovery and Data Mining (WEBKDD '99 Workshop) (SIGKDD '99), pp. 25～30

13 Spiliopoulou M, Faulstich L C, Winkler K. A Data Miner analyzing the Navigational Behaviour of Web Users. In Proc. of the Workshop on Machine learning in User Modeling of the ACAI'99 Int. Conf. Creta, Greece. July 1999

14 Zaiane O, Xin M, Han J. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In: Advances in Digital Libraries conf. Santa Barbara, CA. 1998. 19～29

15 Pitkow J. In search of reliable usage data on the WWW. In: Sixth Intl. World Wide Web Conf. Santa Clara, CA, 1997. 451～463

16 Shahabi C, Zarkesh A, Adibi J, Shah V. Knowledge discovery from users Web-page navigation. In: Workshop on Research Issues in Data Engineering, Birmingham, England, 1997

17 Spiliopoulou M, Faulstich L C. WUM: A Tool for Web Utilization Analysis. In: extended version of Proc. EDBT Workshop WebDB'98, LNCS 1590, Springer Verlag, 1999. 184～203

18 Perkowitz M, Etzioni O. Adaptive web pages: Automatically synthesizing web pages. In submitted to AAAI'98

19 Pitkow J, Bharat K K. Webviz: A tool for world-wide web access log analysis. In: First Intl. WWW Conf. 1994

20 Silberschatz A, Tuzhilin A. What Makes patterns interesting in knowledge discovery systems. IEEE Transactions on Knowledge and Data Eng, 1996, 8(6): 870～974

21 Liu B, Hsu W. Post-Analysis of learned Rules. In: Proc. of the Thirteenth National Conf. on Artificial Intelligence (AAAI'96), 1996. 828～834

22 Liu B, Hsu W, Chen S. Using general impressions to analyze discovered classification rules. In: Third Intl. Conf. on Knowledge Discovery and Data Mining, 1997

23 Padmanabhan B, Tuzhilin A. A belief-driven method for discovering unexpected patterns. In: Fourthj Intl. Conf. on Knowledge Discovery and Data Mining, New York, 1998. 94～100

24 Cooley R, Tan P-N, Srivastava J. Discovery of Interesting Usage Patterns from Web Data. In: Proc of the Web Usage Analysis and User Profiling, Lecture Notes in Computer Science, Vol. 1836, Springer-Verlag, 2000

25 Cooley R, Tan P-N, Srivastava J. WebSIFT: The Web Site Information Filter System (1999). In: Proc. of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), Aug. 1999

26 Joachims T, Freitag D, Mitchell T. WebWatcher: A tour guide for the world wide web. In: the 15th Intl. Conf. on Artificial Intelligence, Nagoya, Japan, 1997

27 Bowman C M, Danzig P B, Hardy D R, Manber U, Schwartz M F. The Harvest Information Discovery and Access System. In: The Second Intl. WWW Conf. (Chicago, IL, 1994), 763～771

(下转第167页)

知，对D中的任一相容故障模式，均也为C中所包含的相容故障模式，即有 $D\subset C$。同样我们可证得 $C\subset A$。因此有 $D\subset C\subset A$。同理可证得 $D\subset B\subset A$。

定理2 B，C，D含义同上述定理1，则B，C，D间满足如下关系：$D=B\cap C$。

证明：当 $a_{ii}=0$ 时，对BGM模型，有 $x_i=0$；对Chwa&Hakimi模型，有 $x_i\oplus x_j=0$，则对 $B\cap C$，有 $x_i=x_j=0$，可知其满足 $x_i+x_j=a_{ij}(a_{ij}=0)$。当 $a_{ij}=1$ 时，对BGM模型与Chwa&Hakimi模型均为 $x_i+x_j=1$，则 $B\cap C$ 为 $x_i+x_j=a_{ij}(a_{ij}=1)$。因此可知，对 $B\cap C$，无论 a_{ij} 取何值时，对 $B\cap C$ 中某一相容故障模式的 x_i 与 x_j 之间，均应满足 $x_i+x_j=a_{ij}$。而对于Maleks模型，D即为满足 $x_i+x_j=a_{ij}$ 的所有相容故障模式的集合。因此有 $D=B\cap C$。

通过定理1与定理2，可以很好地研究四种模型之间相容故障集之间的关系。

定理3 在不存在绝对坏机的情况下，PMC模型与Chwa&Hakimi模型有相同的最优诊断。

证明：由 $C\subset A$ 可知，若某一相容故障模式（假设为p）为Chwa&Hakimi模型的最优诊断，则p至少一定也为PMC模型的相容故障模式。由文[8]中可推知，两故障模型所对应症候分别进行集团运算后，所形成的集团诊断图相同，且均可转化为只含 $u_i+u_j=1$ 项（此时 u_i 与 u_j 均为集团点）。因为不存在绝对坏机的情况，则无集团可确切确定状态。因此由性质3可求出其相同的最优诊断，即PMC模型与Chwa&Hakimi模型有相同的最优诊断。

定理4 设对某一给定的测试图，假设相对于PMC模型其可诊断度为Ta，相对于BGM模型其可诊断度为Tb，相对于Chwa&Hakimi模型其可诊断度为Tc，相对于Maleks模型其可诊断度为Td，则有：$Td\geqslant Tc\geqslant Ta$；$Td\geqslant Tb\geqslant Ta$。

证明：假设此测试图对于PMC为Ta是可诊断的，在此Ta可诊断条件下的解假设为p，则有 $p\in A$，且p为A中含故障点个数最少的相容故障模式。由定理2可知 $C\subset A$。则对C中任一相容故障模式中所含故障点数绝对不会小于p中所含故障点数。可知此测试图对Chwa&Hakimi模型来说，其最优诊断中故障点个数一定不会小于p中所含故障点个数。因此有 $Tc\geqslant Ta$。同样我们可证得 $Td\geqslant Tc$，即有 $Td\geqslant Tc\geqslant Ta$，同理可证得 $Td\geqslant Tb\geqslant Ta$。

定理4说明：1）对一给定的测试图，若其相对于PMC模型为t是可诊断的，则其对Chwa&Hakimi模型一定至少也为t可诊断的，反之则不然。2）为达到相同的t可诊断性，PMC模型所需的测试数不会少于Chwa&Hakimi模型所需的测试数。满足定理4的其他模型之间也有类似关系。

结论 上述性质和定理在对实际应用的网络诊断中有着很好的指导作用。如对虚拟专用网，由于其网点的规模性及分散性，每次诊断将占用大量资源，而当测试模型改变后，由上面的定理可推出一些相容故障模式而不需进行重新的诊断计算；并且在上述定理的指导下，选取适当的测试模型，可有效降低测试边的数目（即降低物理链路的数目），节约企业组网成本，提高诊断效率。又如在无线ad-hoc网络中，由于网路拓扑结构的不断变化性，很难再用图论中某些性质，对诊断带来一定难度，而对于布尔方程，则很容易加入或去除某项，而不需变换上述诊断算法，以高效、实时完成诊断。

本文首次提出了基于布尔方程的系统级故障诊断的概念，并在此基础上推出若干性质与定理，可用于理论上的研究，并对现代不同网络的诊断技术起到一定的指导作用。对于相关性质和对每个模型特殊性质的研究，以及如何将其更有效地运用于网络技术仍在进行之中。

参考文献

1 Preparata F P, Metze G, chien R T. On the connection assignment problem of diagnosable systems. IEEE Trans. Electronic Computers, 1967, 12: 848～854

2 Chwa K Y, Hakimi S L. Scheme for fault-tolerant computing: a comparison of modularly redundant and t-diagnosable system. Information Control, 1981, 49: 212～238

3 Malek M. Undirected graphs models for sytem-level fault diagnosis. In: proc. 7th symp. Comput. Architecture, 1980. 31～35

4 Bianchini R, Buskens R. Implementation of On-Line Distributed System-Level Diagnosis Theory. IEEE Trans. Computers, 1992, 41(5): 616～626

5 Duarte E Jr, Nanya T. A hierarchical adaptive distributed system-level diagnosis algorithm. IEEE Trans. Computers, 1998, 47(1)

6 Jeon G, Cho Y. A system-Level Diagnosis for Internet-based Virtual Private Networks. In: FTCS-29. June 1999

7 Chessa S, Santi P. Comparison-Based System-level Fault Diagnosis in Ad-Hoc Networks. In: Proc. IEEE SRDS 2001

8 Zhang Dafang, Xie Gaogang, Min yinghua. Node grouping in system-level fault diagnosis. J. Comput. Sci. & Technol, 2001, 16(5): 474～479

（上接第132页）

28 Mobasher B, Cooley R, Srivastava J. Automatic personalization based on Web usage mining. In Communications of the ACM, 2000, 43(8)

29 Mobasher B, Cooley R, Srivastava J. Creating adaptive web sites through usage-based clustering of URLs. In: IEEE Knowledge and Data Engineering Workshop (KDEX'99), 1999

30 Luotonen A. The Common Logfile Format. http://www.w3.org/daemon/user/config/logging.html, 199

31 Silberschatz, Tuzhilin A. On subjective measures of interestingness in knowledge discovery. In: U. Fayyad and R. Uthurusamy, eds. Proc. of KDD-95: First International Conference on Knowledge Discovery and Data Mining, Menlo Park, CA, AAAI Press, 1995. 275～281

32 Klemettinen M, Mannila H, Ronkainen P, Toivonen H, Verkamo A I. Finding Interesting rules from Large Sets of discovered association rules. In: Proc. of the Third Intl. Conf. on Information and Knowledge Management, 1994. 401～417

33 Padmanabhan, Tuzhilin A. Unexpectedness as a Measure of Interestingness in Knowledge Discovery. Decision Support Systems, 1999, 27: 308～318

34 Padmanabhan B, Tuzhilin A. On Characterization and Discovery of Minimal Unexpected Patterns In Data Mining Applications. citeseer.nj.nec.com/padmanabhan01characterization.html

35 Liu Bing, Hsu W, Chen Shu, Ma Yiming. Analyzing the Subjective Interestingness of Association Rules. In IEEE Intellgent Systems, 2000, 15(5): 47～55