Application of Data Mining in Supply Chain Management*

Chen, An[1] Liu, Lu[1] Chen, Ning[2] Xia, Guoping[1]
1: Management School, Beijing University of Aeronautics and Astronautics, 100083
2: Mathematics Institute, Academia Sinica, Beijing, 100080;

Abstract: How to deal with the very large database in supply chain management is a very important problem. Mining association rule and sequential patterns from Large Database has been recognized by many researchers in Database systems and many other related management areas recently. In this paper, the necessary of researching on data mining of supply chain is introduced first, then the four stages of the implementation of data mining are given. Nevertheless many previous works were focused on mining association rules in transaction database and sequential patterns at a single concept level, there exists many phenomenon of multiple level sequence patterns in practice. An effective multiple sequence patterns mining algorithm is raised in this paper. A case study is discussed lastly.

KeyWords: Supply Chain Management, Data mining, Knowledge Discovery from Database, Sequential Patterns

供需链管理中的数据采掘

陈 安 刘 鲁 陈 宁* 夏国平
(北京航空航天大学管理学院，100083) (*中国科学院数学研究所，北京 100080)

摘 要： 供需链管理中一个重要的问题是加快处理各类交易数据，由于经济的全球化趋势，一个先进企业或商业机构会不断地获取到来自于自身以及链中其他企业的各类数据，全国乃至世界各地，包括生产、供应、销售等，或者还包括竞争对手的各类数据，如何从这些数据中获取信息，获取信息后怎样采取行动，这需要数据采掘技术来完成信息的获取和分析任务。本文从供需链的各个层次上研究了数据采掘技术，在制造、零售、银行业等部门分析了利用数据采掘技术的可行性和具体措施，最后给出了一个采掘多层次序列模式的算法思想和案例。

关键词： 供需链管理，数据采掘，知识发现，序列模式

供需链[1][8]是深入客户生产、生活过程的相关或自主实体间的动态可重构供求网络，其目标是实现时间、地点、数量、物料、交易对象的优化管理。供需链管理(Supply Chain Management，简记为 SCM)是指供需链中各环节(环境/设备)内部的计划，设计及管理等活动，以及各环节相互之间的协作，物流、信息流、资金流和组织流的管理等。供需链中各环节包括供应商、制造和组装工厂、分销中心、零售商和最终客户等组成的网络[4]。

目前 SCM 研究的主要内容包括供需链中各个环节的管理，具体体现在制造系统的管理、运输和时间表问题、库存决策等；供需链多环节相互间的协作管理，其中包括买方—卖方协作、生产—分销协作、库存—分销协作等。随着企业要处理的信息量的增大，以及因特网的高速发展，使得 SCM 中信息系统的利用成为必然，目前企业采用大多数信息系统是如 MRPII、ERP 等，也有部分软件宣称实现了 SCM，其中包括一些优化措施和决策支持功能。但是目前系统中的供需链管理模块还没有广泛采用数据采掘技术，而企业数据库使用的现状已开始有了这一需求。

数据采掘(Data Mining)是从大量的数据中采掘出隐含的、先前未知的、对决策有潜在价值的知识和规则。这些规则蕴含了数据库中一组对象之间的特定关系，揭示出一些有用的信息，为经营决策、市场策划、金融预测等方面提供依据。

随着信息技术的高速发展，数据库应用的规模和深度不断扩大，已经从单台机器，局域网发展到因特网的全球信息系统。近年来商业条码的推广，企业的信息发布的管理，以及数据采集工具的发展，都提供了巨大规模的数据，在商业管理、政府部门和工业数据处理等领域中应用了数以百万计的数据库。90 年代初全世界共有数据库总量超过 800 万个，信息总

* This paper is supported by Nature Science Foundation (No.79870009) and 863 High Tech Committee Project (No. 9844-007)

量超过 10^{33} 字节。随着社会、经济与科技的发展，决策所需要的数据量也不断增长，数据和数据库的急剧增长仅仅依靠数据库管理系统的查询检索机制和统计学分析方法已远远不能满足现实需要，它迫切要求自动地和智能化地将待处理的数据转化为有用的信息和知识。数据采掘就是为迎合这种要求而产生并迅速发展起来的，可用于开发信息资源的一种新的数据处理技术[3]。许多著名公司(IBM、INFORMIX、Oracle)都投入巨资研究，并出现了一些产品。许多公司把它作为感知市场变化，提高销售利润，从而使自身所在的供需链进行正常运转的重要工具[4]。

一、供需链中应用数据采掘的必要性和可行性

1.1 供需链中进行数据采掘的必要性　供需链管理是指从供应商、制造商到分销商和最终顾客的一条链，由于现代商品社会逐渐由卖方市场向买方市场转化，出现商品过剩的情况，所以对于一个企业来说，必须及时了解市场顾客的需求，并协调这一关系，才能够在竞争中站住，而对顾客的了解必须通过对于从各类渠道所获顾客数据进行处理而得到。这是供需链管理的重要特征之一[6]。

供需链存在着从不同渠道而来的各类数据。供需链管理包括不断地加快处理各类交易数据的能力，由于经济的全球化趋势，一个先进企业或商业机构会不断地获取到来自于自身以及链中其他企业的各类数据，全国乃至世界各地，包括生产、供应、销售等，或者还包括竞争对手的各类数据，如何从这些数据中获取信息，获取信息后怎样采取行动，这需要数据采掘技术来完成信息的获取和分析任务。

首先，采用数据采掘技术是供需链管理动态性的要求，由于供需链上的企业在组织上是变化的，在生产上是受市场趋动的，但是我们往往无法及时地认识到动态性，或者环境的一些变化一方面隐藏在很多不变因素的背后，另一方面隐藏在诸多变化的因素中，这样必须采用去伪存真的方法，而数据采掘的方法正是要达到这个目的。

其次，采用数据采掘技术是供需链管理中发现商机的要求，要形成一个动态联盟，成为虚拟企业中的一员，只是等待机会是不够的，必须在信息的海洋合作去发现机会、利用机会，而机会的发现需要对有关的信息进行收集整理，必须依靠知识发现机制和数据采掘方法。

第三，采用数据采掘技术是供需链管理中建立协作关系的要求，各个企业间的协作关系不是一成不变的，需要时就建立，不需要时就取消，管理这种关系如果只依赖于以往的经验和数据往往不够，这样在更广泛的空间中寻求协作伙伴就是一项重要任务，这项任务的完成有赖于数据采掘技术。

事件序列是数据的一种常见形式，例如，如果将一个顾客在商场的一次购买视为一个事件，则该顾客在一段时间内的若干次购买按照购买时间就形成了一个购买事件序列。事件序列中的知识发现的一个重要问题就是要从多个事件序列中发现经常出现的事件集合。序列采掘是近年来知识发现的一个很活跃的研究领域，它最初是从发现描述一个事件序列的连续生成所遵循的规则开始的。近年来，随着知识发现研究的不断深入和应用的日渐广泛化，事件序列中的知识发现问题正引起越来越多的研究者的兴趣。文[5]给出了序列模式的概念描述，并提出了几种序列模式的采掘算法。

1.2 供需链中进行数据采掘的可行性　在供需链中进行数据采掘是可行的，原因主要在于：数据的持续不断的产生并通过数据库及其辅助工具分类放置；

信息技术是供需链得以存在和发展的重要基础之一，目前信息技术的应用情况可以说是非常广泛，其成本也已大大降低并将继续降低，信息的网上传输加快了供需链各方的交流速度和协调程度，一些用经典方法无法做到的资源的优化配置与利用现在可以通过最大限度地有效利用信息资源来实现[1]。很多企业和组织通过网站发布信息，完全可以利用这些信息组建一个有效的供需链。

目前计算机的计算能力是完全可以应付目前的数据处理，新的软计算算法使得快速计算成为可能；

另外，不断有可用的商业数据库软件、其他专用软件产品开发出来。

数据采掘的支撑技术越来越成熟和完备[9][10]：

1、数据分类技术，包括可用的归纳技术、人工神经网络、基于案例的方法、统计方法

2、统计技术：包括线性回归、逻辑回归、时间序列、检验方法等；

3、分段：概念的聚类、Bayes 聚类等

4、关联分析：关联发现和序列知识发现等

5、偏差分析：包括统计方法和可视化技术。

二、供需链中数据采掘应用的范畴

2.1 制造业中的数据采掘　　在一个供需链中，制造业是比较关键的核心环节，因为它一般和多个供应商、多个分销商发生联系，而其内部又包含了生产、库存等子系统，因而需要处理的数据较多，数据采掘的必要性就越显著。

制造业中的数据采掘一方面除了进行零部件的故障诊断、资源优化、生产过程分析外，还可以对库存的数量控制、库存选址等问题进行决策支持。

库存控制问题在供需链中是一个十分广泛的问题，在制造业、零售行业中都存在，这里我们把制造业的库存控制中采用数据采掘的地方介绍一下：对于一个较大的生产型企业来说，从供商处采购到一定数量的产品并放置到合适的地方进行存储或者把成品放置在合适的仓库中进行存储，如果配置不合理，很容易造成库存分配的混乱，并增加转移库存的费用。一种方法是通过把制造分厂或车间分为几个组所有同组的分厂/车间得到同样份额的产品，另一种方式是通过各分厂/车间生产的历史数据及模式来针对每一个预测未来的需求；第三种是在几个主要的地点来检验库存的效益，并根据生产以及销售状况来决定库存分配，对于这个问题没有一个唯一的正确回答，但是我们可以采用数据采掘的方法可以为更有效的分配库存提供工具。

2.2 零售行业中数据采掘技术的应用　　我们大多都知道这样一个例子，就是关于超级市场的啤酒和婴儿尿布之间关联规则的发现，事实上，零售业的数据采掘主要就是用于销售预测、库存需求、零售点的选择和价格分析。

我们在这里仍然给出一个超市销售的关联规则的说明，实际上，这正是一个完备的供需链的末端。

零售商对顾客的购买事件应用数据采掘进行知识发现并从而采取积极措施进行优化配置：一个数据采掘工具从超市的大量事务数据中分析后发现，顾客购买了牛奶，通常也同时买面包，这样就得出两者之间存在密切关系的一个结论，采用将这两种食品放在同一货架上的措施后，总销售量大大提高了。这里，数据不在只起到了统计的作用，而变成了有用的信息。

日前，信用卡公司的一个收入来源是把它所拥有的顾客消费数据出售给市场调查公司或企业的市场分析部门，后者将零售数据进行分类处理得到有利于自己未来的知识。

Sportlight 是一种主要用于分析超市数据的系统，它可以采用自然语言理解和商用图表显示；Opportunity Explorer 可以用于多种市场情况。

2.3 金融业、服务业采用数据采掘分析顾客流失现象　在供需链中，由于多数的研究是围绕物流和信息流而进行的，往往忽略资金流的有效性，而对于金融业，对其各个层次的顾客数据进行分析，一方面可以采取一定的措施防止顾客的流失，尤其是一些重要顾客；另一方面还可以根据所采集的数据开拓新的市场。这样，伴随着物流和信息流的通畅，资金流也在有效的组织下有序流动。

由于不断增大的竞争压力，西方国家银行业一个重要的课题是考虑如何使用较少的服务成本吸引更多的顾客，防止目前的顾客流失。

Integral Solution 运用神经网络和归纳规则方法开发了一种通过调查顾客访问次数而预测未来情况的发现系统。

三、供需链中进行数据采掘的过程

一般地，供需链中的商业企业通过数据采掘进行优化管理要遵循四个步骤，见下图：

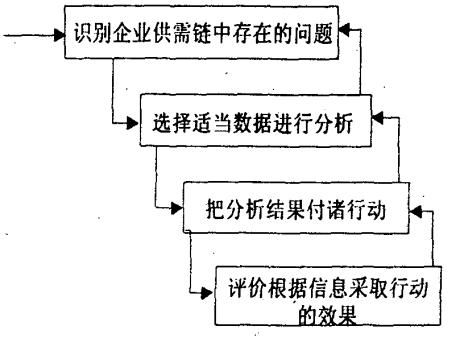

图 1　供需链中通过数据采掘进行优化管理的过程

3.1　根据问题进行分析，找到问题所关联的数据　这也是进行数据采掘所必须的输入，对于制造业来说，它主要有几个过程：新产品的上市计划、定价产品与服务，目标市场定位、理解顾客消减的原因。没有数据采掘，以上的过程是很难实现的。

例如，遇到的问题可以是：

空调的销售情况为什么北京比上海要强？

在一次对数据的一个长期的观察过程中，季节性因素到底有多大的影响力？

在顾客服务上花费更多的钱究竟对保持顾客或吸引新顾客上有什么益处？

3.2 针对问题，对相关数据进行处理　　有了问题，

下一步的工作就是搜集相关数据，产生可以理解的方案，指导进一步的行动。事实上，这一阶段的工作是狭义上理解的数据采掘的过程。

数据是大多数企业流程的核心部分，它是由许多与之相关的基础事务形成，如零售商、通讯、制造、运输、保险、银行等；如果内部数据不足的话，那么外部数据可以增强内部数据(如通过提供零售顾客的人口统计、生活方式、信用信息、金融、市场信息)；当然仅仅根据数据发现模式是不够的，必须做出相应并采取行动，最后将数据转化为信息，信息转化为行动，行动带来价值。这才是数据采掘的中心任务和中级目的。

这其中需要解决一些技术性的问题，比如数据格式的确定，数据库的设计，算法的速度问题等。

首先，数据来源不同，其完备性、唯一性不一定能保证，比如竞争对手的数据就不一定能够搜集完整，而同一个词在不同的企业或不同的产品中的使用意义不一定相同；所以首先需要对数据进行预处理。

最后通过数据给出的可理解的信息一般是符合统计规律的，不一定完全吻合，所以在分析数据给出结果的过程中应该考虑到小概率事件的影响，对不能忽略的小概率事件给予充分的重视。

3.3 给出解决方案，并付诸行动　数据中存在的信息确定后，下一步也是最关键的一步是采取行动。一般来说，信息可能是这种形式的：应该根据顾客需求开发新产品；对顾客的服务电话应该采取优先级服务的原则，不一定完全统一对待；改变组织形式，缩短相应时间；对市场的反应由被动相应而主动行动。

3.4 分阶段评价解决方案所带来的效果，并进一步改进现有方法　采取行动后，一个数据采掘的周期并没有结束，如何评价以上过程的效应是一个很重要的事情。它标志着以上努力的一个反馈信息，提供以上措施是否有成本上的降低或市场与顾客的积极反应。

我们可以采用这样一些标准来衡量：

成本的竞争力与成本控制、库存预测情况四否更准确、质量、服务、运输、新品开发与市场接收程度、订货周期是否缩短等。

四、多层次序列模式的采掘算法与案例分析

本节将给出利用多层次序列模式采掘的快速算法在顾客购买序列中进行数据采掘的案例分析。

在给定的交易数据库中，每个商品称为一个数据项，设 $I=\{i_1,i_2,\ldots,i_m\}$是数据项的集合。每条交易包含以下域：顾客标识号，交易时间，该交易所购买的商品集合。假设不考虑购买商品的数量，且同一时间一位顾客只进行一次交易。数据项集(itemset)是数据项的非空集合，记为$(i_1,i_2,\ldots,i_m)$,其中 i_j 是数据项。D 中每条交易对应一个数据项集。数据项集包含的数据项的个数称为数据项集的长度，长度为 k 的数据项集称为 k 维数据项集(k_itemset)。序列是数据项集的一个有序集，记为$(s_1, s_2,\ldots, s_m)$，其中 s_j 是数据项集。序列包含的数据项集的个数称为序列长度，长度为 k 的序列称为 k 维序列[7]。

数据项集 X 的支持度 Support(X)是一次交易中购买 X 的顾客数量与 D 中总顾客数量之比。序列 A 的支持度 Support(A)是交易数据库 D 中包含 A 的顾客序列的数量与 D 中总顾客数量之比。对给定的第 l 层最小支持阈值 minsup[l]，如果数据项集 X 的支持度 X.support ≥ minsup[l]，则 X 在第 l 层是大数据项集(Large Itemset)。 如果序列 A 的支持度 A.support ≥ minsup[l]，则 A 在第 l 层是大序列(Large Sequence)。大序列实际就是大数据项集的有序表。序列模式是具有最小支持阀值的最大序列。算法的过程为：首先扫描预处理（排序和代码化）后的交易数据库 T[1]，求出第一层的所有大数据项集（即一维大序列），利用一维大数据项集对 T[1]过滤，删除每条交易中的小数据项以及不包含任意一维大数据项集的交易，得到过滤后的交易数据库 T[2]。然后将 T[2]转换成由一维大序列构成的顾客序列，利用 AprioriAll[5]算法求出第一层的所有大序列，然后依次求出第 2,3,...,max_level 层的大序列。若某一层的一维大数据项集为空或者到达最大层，则算法停止。

设顾客数为 4，第一层次最小支持为 Level_1: minsup [1]=50%×4

T[1]

顾客识别号	购买的数据项
1	<(111,211),(111,222)>
2	<(112,411),(111,231)>
3	<(111,211,413)>
4	<(211,323),(323,411,534)>

L[1,1]

序列	支持	映射
<(1**)>	3	1
<(2**)>	4	2
<(1**,2**)>	3	3

T[2]

顾客识别号	顾客序列
1	<(111,211),(111,222)>
2	<(112),(111,231)>
3	<(111,211)>
4	<(211)>

T'

顾客识别号	映射的顾客序列
1	<(1,2,3),(1,2,3)>
2	<(1),(1,2,3)>
3	<(1,2,3)>
4	<(2)>

L[1,2]

序列	支持
<1,2>	2
<1,3>	2

L[1,3]=∅;

LL[1]={<(1**)>,<(2**)><(1**,2**)>,<(1**),(2**)>}

Level_2: minsup [2]=50%×5

L[2,1]

序列	支持	映射
<(11*)>	3	1
<(21*)>	3	2
<(11*,21*)>	2	3

T'

顾客识别号	映射的顾客序列
1	<(1,2,3),(1)>
2	<(1),(1)>
3	<(1,2,3)>
4	<(2)>

L[2,2]= ∅; L[2,2]={<(11*)><(21*)><(11*,21*)>};

Level_3: minsup [3]=50%×5

L[3,1]

序列	支持	映射
<(111)>	3	1
<(211)>	3	2
<(111,211)>	2	3

T'

顾客识别号	映射的顾客序列
1	<(1,2,3),(1)>
2	<(1)>
3	<(1,2,3)>
4	<(2)>

L[3,2]= ∅; LL[3]={<(111)><(211)><(111,211)>}

五、结论

在供需链管理的研究和实践中，序列模式都是广泛存在的一种数据模式，从而这种模式的数据采掘是一个具有实际意义的任务，它能发现大量交易数据中隐含的序列知识。

事实上，在供需链中的各个环节都可以看到数据采掘的应用，因为对于一个竞争的市场经济环境，每一个企业都必须依赖于自身的能力和对市场机遇的把握来谋得生存的空间，这样就使企业数据库和在网上传输的数据进行分析就成为必然[11]。一个制造企业需要对分销商和客户需求进行数据分析，原材料和半成品的供应商对制造商乃至最终客户的信息的有效分析有助于安排生产[12]。

参 考 文 献

1. Chen An, Liu Lu, Li Gang, *Agile Supply Chain Management Based on Agent Technology*, Proceedings of the forth Asia Pacific Decision Sciences Institute Conference, Shanghai, China, June 9-12, 1999, p589-591

2. Chen An, Liu Lu, Xia Guoping, Chen Ning, *Modeling and Analysis Of Production and Distribution System In Supply Chain Management*, in Proceedings of the 26th International Conference on Computer and Industrial Engineering (C&IE), Australia, 1999, p340-345

3. Rakesh Agrawal, Tomasz Imielinski, Arun Swami, *Mining Association Rules Between Sets of Items in Large Databases*; Proc. ACM SIGMOD; May, 1993; p207--216

4. Rakesh Agrawal, Ramakrishnan Srikant, *Fast Algorithm for Mining Association Rules*, Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994, p487--499

5. Chen Ning, Chen An, *Discovery of Multi-Level Sequential Patterns from Large Database*, Proceedings of the International Symposium on Future Software Technology, Software Engineers Association, Nanjing, Oct. 24-30, 1999, p169-174

6. 陈安, 刘鲁, *供应链管理问题的研究现状及挑战*, 系统工程学报, 2000 (将发表)

7. Rakesh Agrawal, R. Srikant, *Mining Sequential Patterns*, Proc. 1995 Conf. Data Engineering, Taipei, Taiwan, March 1995

8. Hau L. Lee, et. al., *Information Distortion in a Supply Chain: The Bullwhip Effect*, Management Science, Vol.43(4), April 1997, p546-558

9. J-S.Park, M-s.Chen, P.S.Yu, *An Effective Hash Bashed Algorithm for Mining Association Rules*, Proc. ACM SIGMOD, May 1995, p175-186

10. David W. Cheung, et.al., *Efficient Mining of Association Rules in Distributed Databases*, IEEE Transactions on Knowledge and Data Engineering, December, 1996, Vol.8(6), p911-922

11. 陈宁, 周龙骧, *Internet上的数据采掘*, 计算机科学, No.7, 1999, p44-49

12. 陈宁, 陈安, *数据采掘的现状与未来*, 自然, 1998 年 5 月, Vol.20(3), p156-160