通讯行业客户行为的关联挖掘

梁 循
(1. 北京大学 计算机科学与技术研究所,北京 100871;
2. 斯坦福大学 管理科学系,美国 加利福尼亚 94305)

摘 要:提出了一种基于关联规则挖掘的聚类方法。首先,通讯行业客户行为的原始数据经过数据预处理转变为地区间的"距离"数据。其次,由于地区是"漂浮"的,不再是"刚体",而是一种抽象的"柔性"距离,使用关联规则进行挖掘成为一种好的选择。文中对通讯行业客户行为进行了基于关联规则的建模,较好地嵌入了关联规则的框架。在数据实验后,提炼出了知识,发现东南亚客户聚成一类,以此为模式,得出了"在南美发展业务是错误的"的结论,该结论在挖掘之前是没有意料到的。实践上,该结论阻止了相应公司的南美发展计划,为公司度过后来的硅谷经济萧条时期省下了上百万美元的"战略储备"资金。

关键词:关联挖掘;聚类;连通子图;客户行为
中图分类号:TP391 **文献标识码**:A **文章编号**:1005-3751(2006)03-0001-04

Mining Association Rules in Customer Behavior in Telecommunication

LIANG Xun
(1. Institute of Computer Science and Technology, Peking University, Beijing 100871, China;
2. Department of Management Science, Stanford University, Palo Alto, CA 94305, USA)

Abstract: Presents a method of clustering based on mining association rules. First, the raw data for the customer behavior in telecommunication are transformed into the "distances" between the areas. Second, since the areas are "floating", as opposed to "rigid", the application of the association rule technique to the "flexible" distances turns out to be an adequate option. The customer behavior in telecommunication is calibrated into the framework of association rules. Experiments hint us the pattern by grouping the customers into a cluster. Based on the pattern, an analysis results in a conclusion that "it is a wrong decision to develop the business in South America", which is unexpected before data mining. In practice, the conclusion was applied and prevented the plan of South America, saving some one million US dollars of the strategic fund for living through the years of stagnant economy in Silicon Valley for the corporation.

Key words: mining association rules; clustering; connected subgraph; customer behavior

0 引 言

数据挖掘自诞生之日起,其使用性就时常受到争议,因为它挖掘出的结果常常是意料之中的东西,挖掘只是起了个验证作用。后来,关联规则找出了"啤酒和婴儿尿布"关联,出乎挖掘之前的预料,从而引起了人们对数据挖掘的兴趣,极大地改变了学者们对数据挖掘,特别是关联规则挖掘的看法。近几年,关联规则研究在数据挖掘研究中成长为一个不小的热点[1,2]。不过,数据挖掘的研究从整体上说,挖掘出类似上述"新颖性"的报道仍然是不多见的。文中给出了一个笔者在实践中得到的、依靠关联挖掘得出意料之外结论的另一个例子,这个例子在实践中已经被采用,并取得了预期经济效益。为了叙述方便,以下将"国家或地区"统称为地区。文中,将一个地区当做一个客户。

在1999~2001年美国硅谷经济过热时期,笔者曾在美国硅谷的一个高科技公司(用K表示)工作。K公司是做全球互联网传真业务的,即客户在传真机上发出的国际传真,实际并没有通过电话线发出,而是进入本地的K公司的服务器,再通过互联网到达目标地区的K公司的服务器,然后再转到目标号码的传真机上,这样相当于打了两个本地电话,其费用和一般比国际电话费少很多。在技术上,需要在拨出号码和目的地号码的地区设置计算机服务器,并且由各计算机服务器之间传递传真信息。

由于互联网的月租金是固定的,且很低廉,所以,K公司的服务报价就比(特别是对亚洲、非洲、南美洲的)电话公司低得多。所以,平时在业务中使用传真多的亚洲公司

收稿日期:2005-09-21
基金项目:留学回国启动基金资助项目(4131522);国家自然科学基金资助项目(70571003)
作者简介:梁 循(1965—),男,北京人,博士,博士后,MBA副教授,研究方向为电子金融、数据挖掘。

就被吸引过来，成为K公司的客户。这些公司主要来自新加坡、泰国、台湾、印度尼西亚、马来西亚等东南亚地区。所以，通过2年多的时间，K公司在上述这些地区的很多城市逐步设置了服务器，并聘当地软件工程师维护。在当时的技术条件下，K公司的这项业务还是很赚钱的，也很有发展前途。

K公司的董事会注意到有不少传真是发向南美的，由于南美没有服务器，K公司只好将这些传真通过互联网统统传到美国，在美国使用AT&T的电话线送到南美(美国的电话费比亚洲低得多)。不过，这样做只好每月付给AT&T大约10万美元左右的电话费(前15个地区见表1，其中第14、15位是位于南美洲的地区)。虽然K公司加给了客户更高的价格，并没有赔本，但如果在南美有服务器，将这部分传真在当地发出去，则会省下不少钱。所以，为了节省这部分开支，K公司的董事会提出了逐步在南美各城市投资设置服务器，扩大经营的计划。

表1 每月付给AT&T费用的前15地区(单位：美元)

1	Japan	$7 735.42
2	Australia	$6 867.85
3	South Korea	$5 987.16
4	USA	$5 604.87
…	…	…
14	Brazil	$1 581.82
15	Argentina	$1 440.61

但是，笔者当时使用了关联数据挖掘的方法，得出相反的分析结论。K公司的董事会根据分析结果，否定了原来投资和开辟南美市场的计划，而集中精力收回在亚洲的投资，避免了上百万美元投资的损失。后来，美国高科技公司出现不景气现象，K公司能够活下来，“否定南美计划”不能不说是一个重要决策。

1 预处理

原始数据为各地区间传真量的二维截面数据，即对某一个时间(月)来说，是一个二维数据组，即可以将各地区的月传真量用表2来表达。

设共有N个地区。目标是将这些地区自动聚类，每类中含有若干个地区。由于每一个客户又向其它$N-1$个客户发数量不等的传真，直接使用这$N-1$个数据来衡量该客户指标太多，观察到客户之间的收发量相差不多，故可以约简这些数据。数据约简是数据挖掘的预处理步骤之一[2,3]。具体地说，就是将收发的量求和，作为两地区的总传真量。这种约简在实践中也是可以解释的，即如果两公司有业务往来，在大多数情况下，不可能只有一个公司向另一个公司单向发传真，必然是有来有往。不难看出，如果写出矩阵，这是一个对角线元素为0的上三角阵，共有元素C_n^2个。将总传真量求倒数，并认为这是两两地区的相互“距离”(见表3)。显见，在这个问题里，地区没有坐标值，只有相互间的“距离”。

表2 各地区的某月传真量(单位：小时)

收/发	Singapore	Thailand	Taiwan	Indonesia	Malaysia	…	Egypt	…	(N)
Singapore		560	1230	1100	1170	…	2.1	…	…
Thailand	1120		910	970	930	…	1.56	…	…
Taiwan	1100	780		600	860	…	0.97	…	…
Indonesia	970	1100	780		1050	…	3.1	…	…
Malaysia	810	1200	1040	890		…	1.81	…	…
…	…	…	…	…	…	…	…	…	…
Egypt	1.75	1.5	1.34	2.1	1.1	…		…	…
…	…	…	…	…	…	…	…	…	…
(N)	…	…	…	…	…	…	…	…	…

表3 各地区的“距离”

	Singapore	Thailand	Taiwan	Indonesia	Malaysia	…	Egypt	…	(N)
Singapore		1/1680	1/2330	1/2070	1/1980	…	1/3.85	…	…
Thailand			1/1690	1/2070	1/2130	…	1/3.06	…	…
Taiwan				1/1380	1/1900	…	1/2.31	…	…
Indonesia					1/1940	…	1/5.2	…	…
Malaysia						…	1/2.91	…	…
…	…	…	…	…	…	…	…	…	…
Egypt						…		…	…
…	…	…	…	…	…	…	…	…	…
(N)	…	…	…	…	…	…	…	…	…

这里的“距离”数据有两个特点。第一，由于缺乏坐标，造成了其度量缺少全面性。从表象上看，这些地区都在“漂浮”(见图1)。第二，这个距离是“柔性”距离。设$d(x_i,x_j)$为地区x_i和x_j之间的“距离”。传统上，一般要求距离满足下面4条公理：

(1) $d(x_i,x_j)=0\Leftrightarrow x_i=x_j$；

(2) $d(x_i,x_j)\geqslant 0$，对一切x_i,x_j成立；

(3) $d(x_i,x_j)=d(x_j,x_i)$，对一切x_i,x_j成立；

(4) $d(x_i,x_j)\leqslant d(x_i,x_j)+d(x_j,x_k)$，对一切$x_i,x_j,x_k$成立。

但是，文中的情况是，$d(x_i,x_j)$和$d(x_i,x_j)+d(x_j,x_k)$之间的关系3种情况都存在，即可能$d(x_i,x_j)<d(x_i,x_j)+d(x_j,x_k)$，也可能$d(x_i,x_j)=d(x_i,x_j)+d(x_j,x_k)$，还可能$d(x_i,x_j)>d(x_i,x_j)+d(x_j,x_k)$。一般说来，文中的问题往往不能满足公理(4)的三角不等式，所以，它们的“距离”不可理解为“刚体”，而是一种抽象的“柔性”距离。但是，为了方便，文中从广义的角度上也称它为距离。

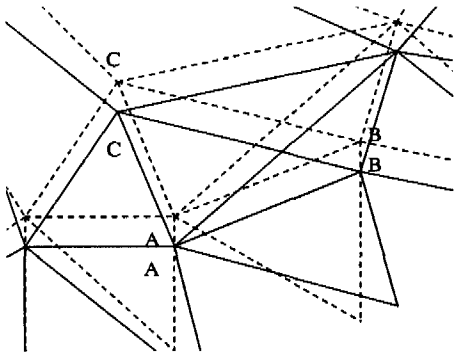

图1　"漂浮"的各个地区及其"距离"

进一步说,由于欧洲、北美洲、南美洲的传真量小,变换成"距离"后,这些洲的地区之间的"距离"就很大,以至于无法在一个平面中画出,只能在一个类似褶皱的"山脊"面的曲面上画出(为了方便,称它为"山脊"曲面)。在这个"山脊"曲面上,东南亚地区集中在其山顶处,而欧洲、北美洲、南美洲的地区在山顶外散开并有褶皱,以使彼此间在"山脊"曲面上有较大的距离。如果将这个"山脊"曲面"压平"为平面,则在山顶外的部分会"皱起来"(见图2)。可以想象,即使本问题有坐标,如果放在欧几里得空间去解决,也很麻烦。下面采用关联规则挖掘的方法,避开了直接在欧几里得空间挖掘的问题。

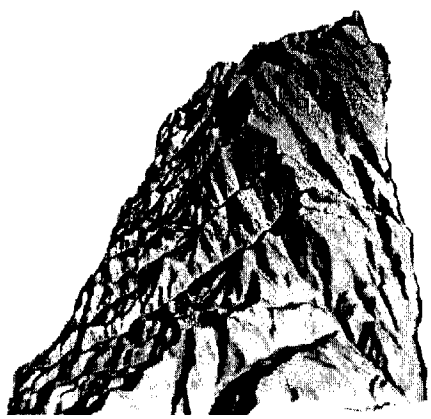

图2　"山脊"曲面上的挖掘

2　顾客行为的关联挖掘建模

显见,由于没有坐标,又不是"刚体",如果要形成类,那么类的质心不好找,半径、直径也不好定义,即基于刚体"距离"的聚类方法直接用在本处并不方便,所以,常见的聚类方法[2]就不适用这种情况。故考虑使用关联规则来解决挖掘问题。

关联规则是一种简单但很实用的数据挖掘方法。关联规则算法本质上是一种对条件概率、联合概率的方法的简化,并在这个简化过程中注意了对数据库扫描次数和效率的改进,从而使算法更加实用化。

在使用关联规则算法时,需要检查 k 个事务项同时出现的情况[1,2,4]。在零售业,这种情况很直观。但是,要想将关联挖掘拓广到其他领域,例如文中研究的通讯领域的客户(即地区)分类情况,需要做进一步的建模分析。

(1)将"距离"作为准事务,形成准事务库和事务库。

考虑上面提到的 C_n^2 个"距离",并将两地区直接的"距离"作为一个准事务,即,对 N 个地区,将这些数据按时间(月)排列起来,形成准事务数据库。对同一个月份,只要 k 个"距离"同时小于给定的阈值 λ,则认为这 k 个"距离"准事务项同时出现了,变成了 k 个"距离"事务;相应地,将所有小于 λ 的"距离"事务组成事务库 D。

接下来,需要检验各"距离"事务集合的并集 $X \cup Y$ 的支持度,如果它大于事先给定的最小支持度阈值 minsup,就认为 $X \cup Y$ 是频繁项集。

(2)"距离"事务的连通性。

在文中的问题中,要使挖掘出的"距离"事务有实际意义,与之相应的地区还必须是直接连通的。也就是说,通过关联规则挖掘出的一个 basket 里的"距离"事务还可能分成不同的簇,簇内连通,簇间非连通。

为了叙述方便,先给出连通的定义。

定义1(图)　对 N 个地区,有 N_0 个"距离"的网状结构,称之为图。

定义2(连通图)　对 N 个地区的图,如果所有地区均能够直接或间接(即通过其它地区)连通的,称之为连通图。

定义3(全连通图、直接连通图)　对 N 个地区的连通图,如果存在 $N_0 = C_n^2$ 个"距离",即所有地区均是连通的,称之为全连通图,或直接连通图。

定义4(子图)　对 N 个地区的图,选出由 N_1 个地区组成的图称为子图。

定义5(非连通图)　如果在图中存在两个子图,它们之间找不到一条直接或间接的一组"距离"来连通,则称该图为非连通图。

定义6(全连通子图、直接连通子图)　如果子图也是直接连通图,则称它为直接连通子图。

不是基于直接连通子图的"距离"集合无法组成一个类。显见,对给定的 λ 和 minsup,基于关联规则的聚类方法找出的"距离"事务频繁项集,不止是一个直接连通子图,即聚成了不止一个类。

3　实验研究

使用了1999年7月至2000年12月共18组数据,其中每1组数据由 $C_{161}^2 = 12880$ 个数据组成,代表着161个地区之间的"距离"。于是,有18组"漂浮"地区的"距离"准事务的数据,共 $12880 \times 18 = 231840$ 个数据。

给定 $\lambda = 1/800$。λ 越大,"出现"的准事务越多;λ 越小,"出现"的准事务越少。给定项集 $X \cup Y$ 的最小支持度阈值 minsup=70%。在问题中,东南亚的地区明显成簇。minsup 过小造成包括的地区太多,minsup 过大又会造成包括的地区太少,两种情况都不好发现关联规则,70%是一个折中的数值,是在实验中不断试出的相对较好值。

最后，Singapore，Thailand，Taiwan，Indonesia，Malaysia等东南亚及其个别周边地区(例如 Japan，Australia)"聚"成一类，而其余地区成为另外一类。南美由于数据不多，观察不出有什么聚类倾向。

在数据挖掘中，发现亚洲客户的传真主要传向亚洲，发向南美的毕竟是少数。即，亚洲人倾向和亚洲人发传真、做生意，或"近邻间做生意的较多"。这里把它当成一个模式(pattern)。依此类推，南美人倾向和南美人发传真、做生意。所以，如果 K 公司在巴西、阿根廷等地的主要城市设置服务器，就会造成发向南美其它城市的传真大量增加，不仅没有节省支出，反而还会增加支出。这种现象会一直持续到在南美各主要城市都建立了服务器为止，这个过程大约需要 2 年的时间，即在 2 年内支出增加，只是换得了南美市场占有率，并获得不了什么实质性的盈利。

应当承认，我们的数据和 K 公司的服务器地区有关系，因为在有服务器的地区，K 公司必然加大广告宣传力度，使该地区的总传真量增大。所以，数据挖掘是"公司级"的，是针对自己公司的业务情况，为该公司服务的。

目前，K 公司已经不再做全球互联网传真业务，因为随着高科技的发展，互联网电话出现，电话费也一降再降，K 公司的业务已经没有盈利空间了。粗算下来，建服务器、设置专人维护、2 年内增加的 AT&T 电话费支出，至少是上百万美元的投资消耗，而南美市场占有率也随着盈利空间的枯竭变得毫无意义。因此，笔者的这项数据挖掘工作，为 K 公司节约了相当大的成本，为 K 公司在接下来的硅谷经济萧条时代留下了具有"战略储备"意义的资金。

4 结束语

事实上，有了"啤酒和婴儿尿布"相关联的结论后，各商家会将这两种商品摆放在一起。不过，由于午夜去超市买"啤酒和婴儿尿布"的单亲父或母的消费额毕竟有限，"啤酒和婴儿尿布"相关联的发现的学术价值比实用价值要大。但是，文中的结果就不同了，因为这个结果不仅和"啤酒和婴儿尿布"一样有学术价值，而且它被应用到实践中，并取得了巨大的经济效益。

从理论上看，文中实质上是使用关联规则，提供一系列问题转换，完成了聚类分析。形象地，可以将转换前的地区看作"原始变量"，而将"距离"看作"衍生变量"。而使用的是"衍生变量"，将其看作事务，并借助建立在与"原始变量"相关的连通子图的概念，技巧性地使用关联规则、无"教师指导"[2,5]地挖掘出了所需要的类。

此外，在理论上，文中处理"漂浮"的空间数据的方法，也对空间挖掘问题有一定的借鉴意义。对"山脊"曲面的处理方法，没有放在欧几里得空间中去解决，而是引用关联规则直接挖掘出了我们想要的东西。

从应用上看，文中的结果还可以推广到通讯领域更多的问题，例如，手机等通讯工具也通常是打向本地的多，即有"近邻"行为。

参考文献：

[1] Han J, Kamber M. Data Mining - Concepts and Techniques [M]. New York: Morgan Kaufmann,2001.

[2] 李雄飞，李 军. 数据挖掘与知识发现[M]. 北京：高等教育出版社,2005.

[3] Hand D, Mannila H, Smyth P. Principles of Data Mining [M]. London: The MIT Press,2001.

[4] 史忠植. 知识发现[M]. 北京：清华大学出版社,2002.

[5] Dunham H. Data Mining - Introductory and Advanced Topics [M]. New Jersey: Prentice Hall,2003.