利用销售数据的商品影响关系挖掘研究

王金龙[1]，徐从富[2]，徐娇芬[2]，骆国靖[2]

(1. 青岛理工大学计算机工程学院 山东 青岛 266033; 2. 浙江大学计算机科学与技术学院 杭州 310027)

【摘要】数据挖掘技术作为一种有效的决策工具正为企业做出科学决策提供依据。该文针对关联规则挖掘商品间相关性的不足，提出了一种新的计算方法利用销售商的商品销售数据挖掘商品之间的相关性及影响关系。该方法根据商品销售数据的变化得到所有商品销售数据的时间序列，然后计算测量序列的相似度，从而确定商品间影响关系。实验证明了该方法的有效性，同时得到了一些有价值的结果，可用于指导具体商业实践。

关 键 词 商品关系; 数据挖掘; 分段线性化; 时间序列
中图分类号 TP311.13 文献标识码 A

Study of Influence Correlation Mining among Commodities Based on Sale Data

WANG Jin-long[1]，XU Cong-fu[2]，XU Jiao-fen[2]，LUO Guo-jing[2]

(1. School of Computer Engineering, Qingdao Technological University Qingdao Shandong 266033;
2. College of Computer Science and Technology, Zhejiang University Hangzhou 310027)

Abstract Data mining can help business enterprise get valuable information from continual accumulated and updated data sources. This paper uses seller's commodity sale database to investigate the correlations among commodities. Especially, aiming to the shortage of association rule algorithm in mining the correlation among commodities, this paper proposes a new algorithm. Based on daily sale data record of commodities, we obtain their sale data time series according to the change of commodities' sales, then compare these time series, measure their distance, and finally get correlations of commodities. Some experiments on real data sets validate the effectiveness of our proposed method. And we obtain some valuable results, which can guide the business application.

Key words commodities correlation; data mining; piece-wise segmentation; time series

数据挖掘作为一种系统地检查和理解大量数据的工具，能有效帮助商业企业从不断积累与更新的数据中提取有价值的信息，为企业作出科学决策提供有效的依据[1-5]。如关联规则[6]、决策树[7-8]、聚类[9-10]等技术已在购物篮分析、市场营销和客户关系管理中得到了广泛应用与发展。

在商业智能分析中，一个非常重要的工作是了解商品间隐藏的影响关系，对其研究有助于营销及其他商业运营活动的高效进行。一些商品的热销会带动一些相关商品的热卖，而一些新商品的投放也会影响到相关商品的销售。分析商品间影响关系有利于生产厂家制定生产计划；有利于为销售商安排进货周期和进货种类；有利于建立新的铺货规划，确定哪些商品上架，指导何时促销，有效刺激顾客的随机购买欲望，增强卖场灵活性等。

一些研究者根据顾客购买记录，运用关联规则分析商品间关系，利用支持度的大小体现商品同时销售的比重，信任度体现商品间依赖关系。然而，因为隐私的原因[11]，这类数据很难获得，一些经销商只有自己所售商品的记录，此时关联规则无法应用。本文提出了一种基于销售商的商品销售记录挖掘商品影响关系的方法，利用商品销售量的变化来分析商品，确定商品间影响关系。

1 时间序列的分段线性化

时间序列是按照时间顺序取得的一系列观测值，广泛存在于各种大型商业、医学、工程和社会科学等领域。通过对时间序列的简化及近似表示可压缩时间序列，换来更小的存储和计算代价。保留时间序列的主要形态，去除细节干扰，有利于提高数据挖掘的效率和准确性等。

本文采用文献[12]所提出的bottom_up分段线性

收稿时间：2007 – 09 – 07
基金项目：国家自然科学基金(60402010)；浙江省自然科学基金(Y105250)
作者简介：王金龙(1979 –)，男，博士，主要从事数据挖掘和商务智能方面的研究.

算法。首先，在取不同分段数N时，该分段法可获得不同的表示精度。N越大，段分得越细，较小的偏差就会导致原来的一段被分割成更小的两段，相当于提高了表示的精度或观察的分辨率；反之，N越小，段就分得越粗，相当于降低了表示的精度或观察的分辨率。其次，线性分段法可支持对同一时间序列在不同分辨率下的观察，可帮助实现对时间序列进行多分辨率(也称多粒度)的数据挖掘分析。此外，该方法还具有较高的滤除噪声和数据抽象能力，对时间序列进行线性化分段处理的同时，相当于对它进行了平滑和数据缩减处理，变换后的数据量将大幅度减少。

本文利用该方法对商品销售数据进行处理，能够尽可能地消除数据噪声造成的干扰。

2　商品影响关系挖掘

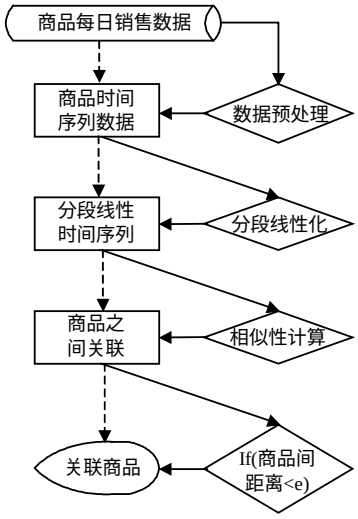


图1　算法总流程图

基于时间序列的线性分段技术可使商品销售数据保持较好的趋势信息，便于商品间相关性度量，以获取商品间影响关系。其算法流程如图1所示，输入每个商品每天的销售量，通过计算，获取商品间影响关系。如ID号为595289的商品可获得如下时间序列：

595289(商品ID):

2005-08-01 42(42为这一天的销售数量)

2005-08-02 52

2005-08-03 45

⋮

针对单个商品的销售数据，本文描述了具体的处理流程。

2.1　数据预处理

分析中，一些商品因销售周期较短影响了结果的准确性，在分析时将其过滤。另外，在选择适当商品后，首先对其销售数据进行预处理，主要是为抹平由大客户或团购因素所引起的大销售量数据。

2.1.1　数据平滑

本文数据平滑处理以14天为时间窗口，取得每个14天销售数据的中位数(根据商业经验，14天的数据在该中位数上下波动)。若中位数小于50，则波动范围不能超过50；若中位数大于50，则波动范围不能超过中位数。波动范围的设定可使数据更平稳。

2.1.2　数据压缩

以天为单位的销售数据有很大浮动性。为进一步对数据进行平滑处理，本文将数据以一周为单位进行压缩，以平滑日销售数据造成的噪声波动，获取更加准确的趋势信息。但以一周为单位的数据压缩会降低大部分商品的振荡。很多商品的销售以天为单位时有很强的规律性，而对于有极大振荡的商品销售数据，其一周销量往往被一两天极高的单日销量所支配。因此在进行时间压缩时，要处理个别孤立点的问题。

2.2　时间序列相似性计算

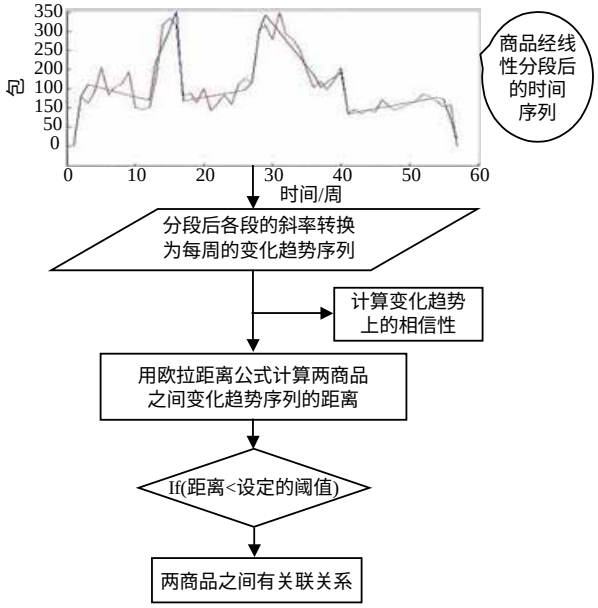


图2　商品间时间序列相似性计算流程图

经过数据预处理，可以进行时间序列的相似性计算以分析商品间关系。因所用销售数据在整体上多数变化平稳，只在局部有大幅度的振荡，所以本文研究的重点转为考虑变化趋势上的相似性，算法流程如图2所示。首先，通过计算所有时间序列分段后各段的斜率来获得变化趋势；然后，对时间序列的变化趋势进行分析，即计算斜率序列的距离，本文采用欧拉距离；最后，基于该度量值及阈值ε对商

品进行相互距离测量，若距离小于阈值ε，则表明两个商品销售间具有正影响关系。

3 实验结果和分析

3.1 数据源简介

本文所用实验数据源自宝洁公司在某大型超市所有门店2005-08-01～2006-07-31(共396天)的销售数据，数据格式包括如下属性：日期、子类号、商店码、条码、商品名称、商品规格、单位、销售数量和金额。商品品牌主要包括：佳洁士、海飞丝、飘柔、沙宣、潘婷、伊卡露、激爽、汰渍、碧浪、玉兰油、帮宝适、护舒宝。商品类别主要包括：牙膏、洗发露、沐浴露、香皂、婴儿纸尿裤、卫生巾、卫生护垫、洗衣粉、洗手液和净肤棉等。

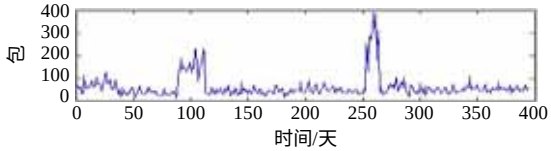
3.2 数据预处理

商品销售数据记录了超市每天出货量，有些突然变大的数据由大客户或团购产生，这些数据不能体现商品的正常销售规律；还有一些负值销售数据是商品未经编码前的记录，应予以去除或对其替换。另外，商品销售数据在日期上不连续，中间会有几天无销售记录，或是值为0，其主要原因为有些商品在开始一段没有销售，上市后这些商品的销售记录从上市当天开始计算，而前面没有记录数据。商品在其生命周期最后一两周的销售数据往往很小，也不能反映商品一般销售水平，应该去除。

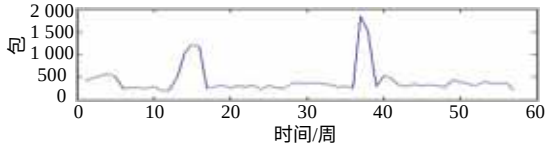
实验共选取了1 000种商品，包括两部分商品：一是所有销售了30天以上的特殊商品(优惠装、促销装等)，共计275个；另一种是生命周期最长的725个正常商品。在分析时，对于相同品牌和类别的商品，若规格不同，认为是不同的商品。在商品整个生命期中，销售记录在日期上并不连续，中间存在许多空缺记录，空缺日期邻近销售量有大有小，没有特殊规律。本文根据空缺天数和实际有值天数的比值作为一个数据好坏的评判标准(值越小越好)。最终，从1 000种商品中挑选了500个商品作为最终的研究对象。

3.3 数据压缩

以一周为单位将数据压缩，结果如图3所示。



a. 以天为周期的销售量

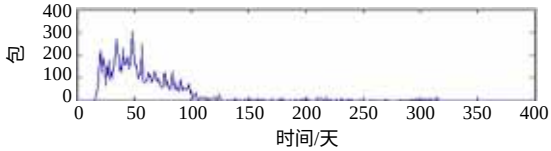


b. 以周为周期的销售量

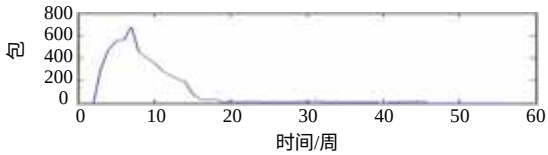
图3 护舒宝瞬洁丝薄日用护翼卫生巾

初步观察所有商品的时间序列，可发现下面的规律：

(1) 促销商品生命周期较短，通常是一个突然的上升和下降趋势，如图4所示。



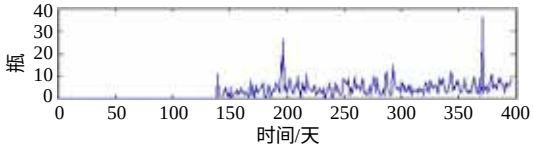
a. 以天为周期的销售量



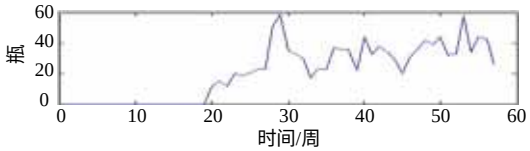
b. 以周为周期的销售量

图4 护舒宝瞬洁丝薄卫生巾(促销装)

(2) 一些商品若型号不同，而名称相同或相近，则它们的销售趋势可能相同。如图5和图6表示名称相同而规格不同的商品销售变化趋势。

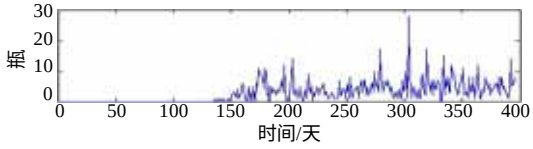


a. 以天为周期的销售量

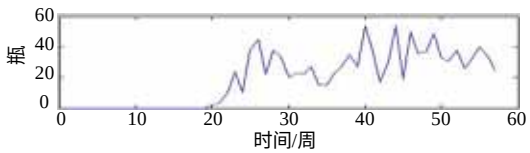


b. 以周为周期的销售量

图5 潘婷乌黑莹亮洗发露200 ml/瓶



a．以天为周期的销售量



b．以周为周期的销售量

图6 潘婷乌黑莹亮洗发露400 ml/瓶

(3) 一些有联系的商品变化趋势相似。初步观测

的结果有利于进行下一步研究的分析验证。

3.4　时间序列的分段线性

对得到的商品销售时间序列进行分段线性，能够将数据点密集的长序列转换成相对稀疏的序列，可达到保留时间序列主要形态，去除细节干扰的目的。经过分段线性的结果如图7所示。

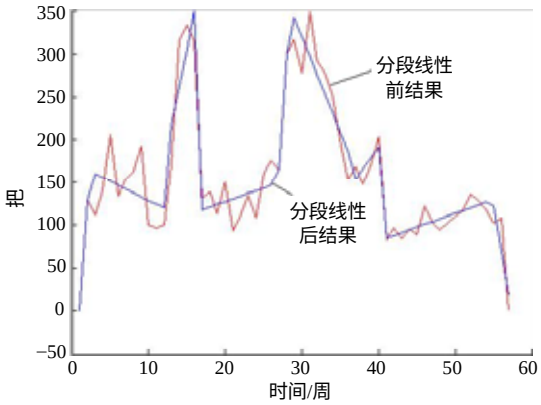


图7　佳洁士三重护理牙刷周销售数据分段线性结果

3.5　商品间相似性分析结果

表1列出了商品之间距离，该结果与观测结果相符。除了容易看出的商品之间相关性外(前3个结果)，在实验中还得到了大量有用信息(后2个结果)，证实了商品之间存在影响关系，同时也验证了算法的准确性。

表1　商品间距离

商品1名称	商品2名称	商品间距离
潘婷乌黑莹亮洗发露 200 ml/瓶	潘婷乌黑莹亮洗发露 400 ml/瓶	17.662 670
护舒宝持久透气无香型超薄卫生护垫	护舒宝持久干爽无香型超薄吸收护垫	19.266 830
潘婷毛磷护理防毛燥润发膜	潘婷锁水亮泽发妆滋养水	2.771 920
伊卡璐草本精华深层修复润发精华素	沙宣垂坠质感润发精华露	8.949 750
飘柔高纯度焗油精华润发精华素	飘柔温泉头皮护理洗发露	19.927 922

4　结　论

本文针对商品销售数据，提出了一种新的挖掘商品之间影响关系的算法。挖掘出的商品影响关系可用于指导销售商、生产厂家以及超市管理者等的经营决策。

参　考　文　献

[1] BERRY M, LINOFF G. Data mining techniques for marketing, sales, and customer relationship management [M]. 2nd ed. [S.l.]: John Wiley & Sons, Inc, 2004.

[2] APTE C, LIU B, P. D. PEDNAULT E, et al. Business applications of data mining[J]. Communications of the ACM, 2002, 45(8): 49-53.

[3] KOHAVI R, ROTHLEDER N, SIMOUDIS E. Emerging trends in business analytics[J]. Communications of the ACM, 2002, 45(8): 45-48.

[4] 韩家炜, 坎　伯. 数据挖掘概念与技术[M]. 第2版. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2006.

[5] 王金龙. 数据挖掘研究进展[J]. 青岛理工大学学报, 2007, 28(4): 85-88.

[6] AGRAWAL R, SRIKANT R. Fast algorithm for mining association rules in large databases[C]// In Proc. 1994 Int. Conf. Very Large Data Bases (VLDB'94). Santiago de Chile: Morgan Kaufmann Publishers Inc, 1994, 487-499.

[7] QUINLAN J R. C4.5: Programs for machine learning[M]. [S.l.]: Morgan Kaufmann, 1993.

[8] 张　靖, 姚　珍, 唐雪飞. 基于决策树的不完整数据的处理[J]. 电子科技大学学报, 2007, 36(1): 116-118.

[9] PARSONS L, HAQUE E, LIU H. Subspace clustering for high dimensional data: a review[J]. SIGKDD Explorations, 2004, 6: 90-105.

[10] 耿　技, 印　鉴. 改进的共享型最近邻居聚类算法[J]. 电子科技大学学报, 2006, 35(1): 70-72.

[11] VAIDYA J, CLIFTON C. Privacy-preserving data mining: why, how, and when[J]. IEEE Security and Privacy, 2004, 2(6): 19-27.

[12] KEOGH E, CHU S, HART D, et al. An online algorithm for segmenting time series[C]//In Proc. 2001 IEEE Int Conf Data Mining(ICDM'01). San Jose: IEEE Press, 2001, 289-296.

编　辑　张　俊