多维关联规则挖掘在彩铃推荐中的应用①

Application of Multidimensional Association Rule Mining in CRBT Recommendation

管 乐 王 纯 (北京邮电大学网络与交换技术国家重点实验室 北京 100876)

(东信北邮信息技术有限公司 北京 100191)

摘 要: 随着彩铃业务的迅猛发展，彩铃精细化营销的需求日渐突出，单维关联规则难以满足新上线铃音、重点铃音的推荐需求。本文提出了基于数据立方体的多维关联规则挖掘算法在彩铃推荐方面的具体应用方法，有效解决了新上线铃音的推荐问题；并在彩铃推荐过程中提出了铃音的“推荐权重”概念，解决了重点铃音的重点推荐问题。

关键词: 多维关联规则 数据立方体 彩铃 精细化营销 数据挖掘

1 引言

彩铃[1]是目前发展颇为成熟的一项增值业务，受到运营商和用户的青睐。中国移动于 2003 年 5 月推出了“彩铃”业务。此后便得到了市场的热烈响应，彩铃业务迅猛增长[2]。目前，在彩铃的推广方面，各运营商已提出由粗放式营销转向精细化营销的需求，从而促进了数据挖掘—尤其是关联规则推荐算法[3]—在彩铃推荐方面的应用。而目前流行音乐更新迅速，彩铃的流行期相对较短，彩铃的推广时机转瞬即逝。因此，各运营商在彩铃—尤其是新上线铃音—的推广方面，提出了更高的时效性要求。但是目前在数据挖掘算法方面，各运营商采用的多是传统的“铃音-铃音”的单维关联规则算法，由于该算法必须要求铃音具有一定的历史下载记录后才能发现相关规则并进行推荐，故在新上线铃音的推荐方面存在较大困难。另外，对于重点铃音的重点推荐，也是运营商提出的铃音推荐需求之一。

根据以上需求，本文基于多维关联规则挖掘算法，提出了根据铃音分类进行彩铃推荐的思路，有效解决了新上线铃音的推广问题；另外，本文提出了铃音“推荐权重”概念，有效地解决了对重点铃音的重点推荐问题。

2 基于数据立方体的多维关联规则

2.1 数据立方体的生成

在数据仓库中，多维数据模型将数据按数据立方体[4](data cube)的形式组织。数据立方体由维和事实定义。维是指观察数据的角度，维的取值称为维成员。数据的特定角度(维)上还存在细节程度不同的多个描述，这种描述称为维层次。事实表中包含相关维表的关键字和事实的度量。

根据用户给定的挖掘任务，从数据仓库中生成数据立方体，可在此数据立方体上进行关联规则挖掘。而且由于数据仓库中数据立方体被事先全部或部分物化(materialization)存储，从而为关联规则挖掘节省了大量挖掘时间，提高了挖掘效率。

给定一个关联规则挖掘任务，其内容涉及 d_1，...，d_n 个维，并根据用户挖掘需求确定各维的维层次，然后从数据仓库中生成数据立方体。其中每一维包含 $|d_i|+1$ 个数值，$|d_i|$是第 i 维包含的互不相同的维成员个数。在 d_i 维中，前$|d_i|$ 行各代表 d_i 中一个互不相同的维成员。最后一行存储了一个称之为“SUM”的维成员，其中记录了它所对应的维的合计值，这种合计值极大地方便了关联规则的挖掘中支持度的计算。立方体的方格中记录的是对应维成员的频繁度量值，记

① 基金项目:国家杰出青年科学基金项目(60525110);国家 973 计划项目(2007CB307100,2007CB307103);电子信息产业发展基金项目
收稿时间:2008-09-16

为 count。这样涉及 d_1，…，d_n 维数据的一个关联规则挖掘任务就对应一个 n 维的数据立方体 Cube（d_1，…，d_n | count），其中 d_1，…，d_n 是立方体的维，count 是立方体的事实度量。

2.2 关联规则

设 $I=\{i_1, i_2, \cdots\cdots, i_m\}$是 m 个不同项目的集合。D 是所有事务的集合(即事务数据库)，每个事务 T 是一些项目的集合，T 包含在 I 中，即 $T \subseteq I$，并且每个事务可以用唯一的标识符 Tid 来标识。

定义 1. 设 X 为 I 中某些项目的集合，简称为项集（itemset），如果 $X \subseteq T$，则称事务 T 包含 X。关联规则表示为: $X \Rightarrow Y$ 的蕴涵式，这里 $X \subset I$，$Y \subset I$，并且 $X \cap Y = \Phi$。D 中的规则 $X \Rightarrow Y$ 是由支持度(support)和置信度(confidence)来约束的。支持度表示规则出现的频度，置信度表示规则的强度。具体描述是:

$$Support(X \Rightarrow Y) = P(X \cup Y) \quad (1)$$

$$confidence(X \Rightarrow Y) = P(Y|X) = P(X \cup Y)/P(X) \quad (2)$$

同时满足最小支持度阈值（minsup）和最小置信度阈值(minconf)的规则才是有意义的规则，称作强规则。关联规则挖掘的目标就找出数据集中的所有强规则。对于项集 X，如果 X 中包含有 k 个项目，则 X 称为 k-项集。若项集 X 的支持度不小于最小支持度，则称 X 为频繁项目集[3]。

2.3 多维关联规则

传统关联规则挖掘是通过对事务数据库中项目集的重复统计求出频繁项目集，项目集是从一维属性中得到，所生成的是单维关联规则。包含两个或更多个谓词的关联规则称为多维关联规则[5]，如:

Age(X,”19~24”)∧occupation(X,”student”)
buys(X,”laptop”)

如果规则中的谓词只出现一次，则是无重复谓词，被称为维内关联规则；含有重复谓词的关联规则被称为混合维关联规则。本文研究的是维内关联规则。

3 多维关联规则算法在彩铃推荐中应用

3.1 彩铃定制立方体的生成

以彩铃定制记录作为生成立方体的主要依据。具体说来，是将彩铃定制记录与各相关表进行关联，并统计各维度组合后的统计值。将彩铃定制记录维度分为三类，分别为用户维、操作维、内容维，各维度由多个子维度组成，详见。

以彩铃定制记录作为生成立方体的主要依据。具体说来，是将彩铃定制记录与各相关表进行关联，并统计各维度组合后的统计值。将彩铃定制记录维度分为三类，分别为用户维、操作维、内容维，各维度由多个子维度组成，详见图 1。

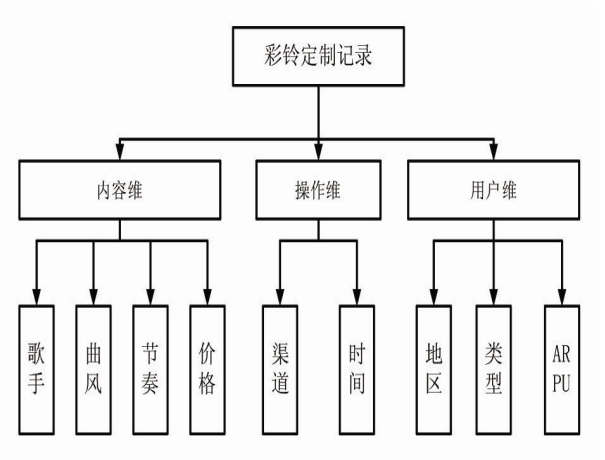

图 1　彩铃定制立方体相关维度

实际生成过程中，以上述各子维度作为立方体维度，通过 SYBASE IQ 数据库服务器提供的立方体生成功能，可以高效灵活地完成立方体的生成工作。SYBASE IQ 是专门面向数据仓库环境的关系型数据库，其独有的按列存储机制及多样的索引机制，保证了在海量数据情况下生成数据立方体的高效性。采用以下 SQL 语句，可以实现立方体的生成:

```
SELECT 歌手，曲风，……，COUNT(*)AS 次数
FROM 彩铃定制记录表
GROUP BY 歌手，曲风，……
WITH CUBE
```

其生成结果如表 1 所示。注意其中的“NULL”值的意义同 2.1 节中的“SUM”，代表本维的合计值。

表 1　彩铃定制立方体

歌手	曲风	……	次数
S1	T1	……	3860
S2	T1	……	6277
……	……	……	……
NULL	T1	……	35332
⋮	⋮	⋮	⋮
NULL	NULL	……	1432352

3.2 关联规则生成

与单维关联规则挖掘不同，在多维关联规则的挖掘中，不是搜索频繁项集，而是搜索频繁谓词集。因此，多维关联规则的挖掘主要分为以下两个步骤：

1）在生成的数据立方体上挖掘满足最小支持度的频繁谓词集；

2）在频繁谓词集中生成用户感兴趣的关联规则。

基于 Apriori 的关联规则挖掘算法[6]的主要工作在于频繁项集的查找，需要利用“一个频繁项集中任一子集也是频繁项集”的性质，利用层次顺序搜索的循环方法来完成频繁项集的挖掘，这一循环方法就是利用 k-项集来产生(k+1)-项集。每挖掘一层需要扫描整个数据库一遍。

在基于数据立方体的多维关联规则挖掘中，也可以采用类似的方法，但此方法的效率较低。实际上，由于数据立方体中已经提供了各层次上的统计值，因此只需比较“次数”列是否不小于最小支持度与定制总次数的乘积，即可得到所有频繁谓词集。这种方法被称为“N 维立方体搜索”(n-D cube search)算法。通过性能研究，由于此算法的时间成本不会随频繁谓词集的复杂程度提高，具有较好的可扩展性，因此，在已经完成数据立方体准备的情况下，推荐使用此算法[5]。

求出频繁谓词集后，提取关联规则就相对容易了，产生过程分为两步：

1）对于每个频繁谓词集 l，求出其所有非空子集

2）对于频繁谓词集 l 的每个非空子集 s，计算规则 $s \Rightarrow (l-s)$的可信度，若 $confidence \geqslant minconf$，则产生关联规则 $s \Rightarrow (l-s)$[6]。

3.3 推荐算法

挖掘出的关联规则被存储在多维关联规则表中，其表结构如图 2 所示。其中以“前”为前缀字段表示规则中的前项部分，以“后”为前缀字段表示规则中的后项部分。规则涉及的部分填对应维度值的编号，规则不涉及的字段置空。

将规则存储后，即可根据用户相关信息找出相适应的规则进行个性化推荐。以针对用户的个性化铃音推荐为例，具体步骤如下：

规则表	
规则编号	Serial
前_歌手	Integer
前_曲风	Integer
.....	Integer
后_歌手	Integer
后_曲风	Integer
.....	Integer
置信度	Decimal

图 2　规则表结构

1）根据某用户的彩铃订制历史记录，与规则表前项部分进行关联，找出所有相适应(即满足所有前项规则要求)的规则，按规则后项中的内容子维度进行分类，对置信度求和，得到每类后项内容子维度的置信度 C_a；

2）将后项内容子维度与所有铃音对应的内容子维度连接，按铃音进行分类，对置信度求和，得到每首铃音的置信度 C_r。

3）每首铃音的置信度与其“推荐权重” μ 相乘，得到每首铃音的最终推荐度 R，即

$$R = \mu C_r \tag{3}$$

4）按照推荐度 R 对铃音进行降序排列，取排名前 n 首的铃音，即为对用户的推荐铃音。

说明：在第 2)步中，通过规则与铃音分类(即内容子维度组合)—而非与具体铃音—的关联，实现了对于新上线铃音的推荐(因为新上线铃音的分类是存在定制记录的，可以产生相关的关联规则)。在第 3)步中，通过对铃音置信度加权“推荐权重”系数，实现了对于重点铃音的重点推荐功能。“推荐权重”可由用户根据需求自行调整，默认为 1。当重点铃音的推荐权重大于 1 时，可以提高其最终推荐度，从而达到重点推荐的目的。

4　结束语

本文根据彩铃应用中的具体需求，提出了基于数据立方体的多维关联规则挖掘算法在彩铃推荐方面的具体应用方法，并在彩铃推荐过程中提出了铃音的“推荐权重”概念，有效解决了新上线铃音推荐与重点铃音的重点推荐问题。经过在某省移动运营商试用，取

(下转第 174 页)

(上接第 157 页)

得良好效果。本文提出的方法不仅适用于彩铃推荐，也适用于无线音乐、多媒体回铃音[7]等相关产品推荐，在其它产品的推荐方面也具有一定参考价值。

参考文献

1 沈奇威,廖建新,王纯,朱晓民.彩铃业务的研究和设计.第九届全国青年通信学术会议论文集.重庆, 2004: 484－489.

2 廖建新.移动增值业务发展趋势.电信工程技术与标准化, 2004,17(5):1－5.

3 Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Databases. Proc 1993 ACM-SIGMOD Int. Conf. Management of Data. Washington D.C, 1993:207－216.

4 Gray J, Bosworth A, Layman A, Pirahesh H. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Total. Proceedings of the 12th International Conference on Data Engineering, New Orleans: IEEE Computer Society, 1996:152－159.

5 Kamber M, Han J, Chang J. Metarule-guided Mining of Multi-dimensional Association Rules Using Data Cubes. Proc. of 1997 Int. Conf. Knowledge Discovery and Data Mining(KDD'97). Aug. 1997: 207－210.

6 Agrawal R, Srikant R. Fast Algorithms for Mining Association Roles. Proc. of 1994 Int. Conf. Very Large Data Bases(VLDB'94). Santiago, Chile, 1994:487-499.

7 沈奇威,廖建新,王纯,蔡斌.多媒体回铃音业务研究.计算机工程, 2006,32(18):231－233.