基于关联分析的粗粒度级个性化信息挖掘*)

Using Association to Mine the Thick-Scale E-commerce Personalized Service Information

李常青　唐世渭　李红燕

(北京大学信息科学中心视觉与听觉信息处理国家重点实验室　北京100871)

Abstract　The data mining technology is used here to provide the thick-scale personalized service. In fact, the building of user's personal information space can lead to classfied indexes which will provide the favorable personalized service with greatly reducing the computing of DM algorithm.

Keywords　E-business, Personalized service, Data mining, Scale

1　引言

个性化服务是指针对用户个人需求的不同，采取不同服务策略的一种个性化服务模式。个性化的健康服务体现在商品信息个性化和配送个性化两个方面。个性化信息服务可以在两种粒度上进行：

(1)细粒度的个性化服务。其主要为网站的注册用户或会员提供。Web 站点上存放着关于这些用户的比较详细的信息，例如性别、年龄、收入范围、健康需求以及定购商品的记录等。在这些信息上进行算法设计，Web 站点能够为这些用户提供针对个人的个性化服务；

(2)粗粒度的个性化服务。它将以用户群为单位来提供相应的个性化服务策略。而根据用户种类的不同，又可分为两个级别来实现。a)对一般用户，这是一种最大粒度的粗粒度个性化服务，只能根据所有数据计算出的商品相关性，提供相关产品，如：当某用户购买了助听器后，系统可提醒用户购买电池；这一层次的个性化服务又可称为广粒度级个性化服务；b)对注册为会员，且有个人信息保存在网站的用户，此时可根据不同的用户群来挖掘商品相关信息，如：当某用户购买了助听器后，系统可提醒用户购买合适价格、品牌的电池。这一层次的个性化服务又可称为中粒度级个性化服务。

本文基于北京大学视觉与听觉信息处理国家重点实验室开发的 PeopleFirst 电子商务原型系统，该系统主要面向全球华人提供医疗信息服务和医疗电子商务服务。文章将从粗粒度个性化服务方面针对 PeopleFirst 系统展开一定的研究。细粒度的个性化服务研究请参考文[1]。

2　问题描述

文[1]对 PeopleFirst 电子商务系统的个性化服务算法作了论述。该算法解决了根据用户个人信息、健康需求和交易行为动态提供个性化服务的策略，服务的粒度是针对每一个用户的。但是该算法在提供粗粒度的个性化服务方面显得有些不足，比如提供联系不太明显的相关产品时。

针对以上不足，本文将采用数据挖掘的方法来进一步完善 PeopleFirst 的个性化服务策略。利用数据挖掘的方法，可挖掘出这样的信息，即：当某用户购买了助听器后，系统会提醒该用户购买电池，这样每一个用户购买了不同的商品后，系统总会根据该用户的购买情况提供不同的相关商品。这就是一种个性化服务，进而可以将用户分成不同的用户群，首先可以将用户分为两类：即一般用户和会员，而针对会员，又可将其分为男性、老年人、月收入3000～5000元，要治疗高血压；女性、年轻人、月收入2000～3000元、有养颜美容需求等等，这样就可实现一般用户这一根据所有数据信息进行相关商品提供的个性化服务(数据没有针对性，不太准确)和针对会员下面的每一小类用户进行个性化服务(相关产品及产品性质提供更为准确)。见图1。

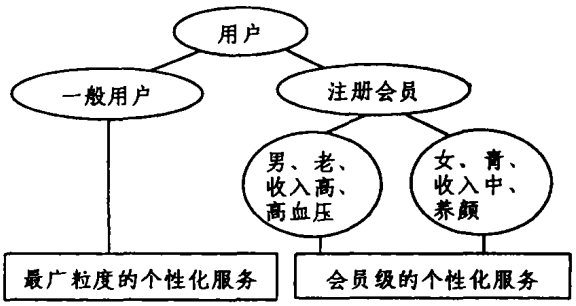

图1　粗粒度个性化服务分类层次

关于数据挖掘的方法，许多著作[2,3]和文[4]都作过论述。本文所采用的主要挖掘方法基于关联规则的采掘。

3　PeopleFirst 电子商务粗粒度个性化信息的知识发现过程

PeopleFirst 电子商务粗粒度个性化信息的知识发现过程采取如下一些步骤：

(1)领域背景及源数据

PeopleFirst 系统主要面向医疗领域提供医疗信息服务和电子商务交易服务。本文所述的个性化服务主要针对电子商务，所以一切与电子商务交易有关的数据都可作为电子商务粗粒度个性化服务研究的源数据。

(2)从源数据中选择目标数据

源数据中包括了商品的分类信息、商品本身属性信息、厂家信息、用户交易信息以及用户个人信息，从广义上来说，所有这些信息对个性化信息的挖掘都是有用的。比如可根据用户注册的个性化信息将用户进行分类，针对不同的用户群，根

*)本文研究得到国家重点基础研究发展规划(九七三)(G199032705)资助。李常青　硕士生，主要研究领域为电子商务、智能信息系统。唐世渭　教授，博士生导师，主要研究领域为智能信息系统，数据库。李红燕　博士后，主要研究领域为数据库，电子商务。

据用户的所有交易信息用关联分析方法来提供产品相关性分析，然后根据用户的个人信息不同，提供符合该用户的最为合适的商品，利用商品的分类信息并利用广义关联规则挖掘算法，可以实现不同概念层的个性化服务。这样用到的信息库有分类信息、商品本身属性信息、用户交易信息以及用户个人信息，至于厂家信息，则可用于进货。但考虑到数据挖掘算法的效率以及更准确地提供个性化服务，数据量要精简和选择。根据这一要求，个性化服务的数据挖掘目标数据为商品分类信息、商品属性信息、用户交易信息以及用户个人信息。

(3)目标数据预处理，去除错误和冗余数据

目标数据选定以后，可以看出其中还有很多的冗余字段。比如用户个人信息包括用户注册名、密码、真实姓名以及性别、商品年龄、收入范围、健康需求等，但真正对个性化信息的挖掘有用的将是后面的四项，这样注册名、密码、真实姓名等个人信息就可以过滤出去。用类似的方法，可将目标信息库中的各种信息进行信息预处理。处理后的结果如下：

a. 商品分类信息：分类编号、分类名称。

b. 商品属性信息：商品编号、商品名称、分类编号、单价、简要说明。

c. 用户交易信息：订单编号、商品编号、会员编号、订货日期、成交额。

d. 用户个人信息：性别、年龄、收入范围、健康需求。

因为年龄等个人信息在注册时已进行了离散化处理，即将年龄离散化为年龄段0～1岁，1～5岁，5～12岁，12～20岁，20～30岁，30～50岁，50岁以上等等，所以不再需要进行数据转换的工作。

(4)进行数据挖掘，发现模式并表达成易于理解的形式

完成目标数据的预处理后，即可进行关联模式的挖掘，建立起诸如这样的模式“80%购买了助听器的顾客也购买了电池”。当把所有这样的关联模式挖掘出来后，将给个性化服务提供巨大的效益。而这样的模式，如：“85%购买了 A 商品，B 商品，C 商品…的顾客。也会购买 Z 商品”，前提尽量的多伴随结果尽量的少，这样的模式将为顾客提供更为准确的个性化服务，同时避免相关商品信息提供的泛滥。

(5)评价和解释发现的模式，必要时反复执行(1)到(5)

关联模式已经建立，评价和解释模式的正确性可从用户对这些相关信息的反应所积累的数据来检验。必要时重新进行(1)至(5)，重新挖掘或改进现有模式。

4 电子商务个性化信息关联挖掘及挖掘工具(ECPI Miner)设计

4.1 广粒度级个性化服务

关联规则采掘在数据挖掘的研究中进行得比较深入。文[5～7]介绍了关联规则采掘的情况，关联规则及其发现算法是1993年由 Agrawal、Imielinski 和 Swami 提出的，后来 Agrawal[8]对算法作了改进，提出了快速发现关联规则算法。描述如下：

设 $I=\{i_1,i_2,\cdots,i_m\}$ 是 m 个不同项目的集合，D 是针对 I 的交易的集合，每一笔交易包含若干项目 $i_1,i_2,\cdots,i_m\subset I$。关联规则表示为 $X\Rightarrow Y$，其中 $X,Y\subset I$，并且 $X\cap Y=\phi$。X 称作规则的前提，Y 是结果。如果 D 中包含 X 的事件有 c%也包含 Y，则称规则 $X\Rightarrow Y$ 在 D 中的可信度(confidence)为 c。如果 D 中有 s%的事件包含 $X\cup Y$，则称 $X\Rightarrow Y$ 在 D 中的支持度(support)为 s。c 和 s 的公式定义如下：

$$c=\frac{\text{同时购买商品 X 和 Y 的交易数}}{\text{购买了商品 X 的交易数}}*100$$

$$s=\frac{\text{同时购买商品 X 和 Y 的交易数}}{\text{总交易数}}*100$$

挖掘关联规则就是对给定的一组交易 D，找出其中所有满足最小支持度(minsup)和最小可信度(minconf)的关联规则。可分为以下两个步骤：

(1)产生所有 support 大于 minsup 的数据项组合，这些组合称为大数据项集；

(2)对于每个大数据项集，产生所有比 minconf 大的规则。

第一步的主要工作思想 Apriori 算法是由种子大数据项集(L_1)开始。对于 k－1大数据项集 L_{k-1}，如果它们的前 k－2项相同，第 k－1项一个比另一个小，则将相同的前 k－2项和不同的最后两个第 k－1项合并，组成 k 数据项集，计算 k 数据项集中每个数据项的支持度，去掉支持度小于 minsup 的数据项集，形成所有 k 大数据项集。如此反复，直到找到所有大数据项集为止。

第二步的主要工作思想 Rule-Generation 算法是对一个大数据项集首先产生后项只有一项的所有规则，再利用 Apriori-gen 产生后项有两项的规则，如此不断反复产生具有更多后项的规则。

广粒度级的个性化服务主要针对未在网站注册任何个性化信息的一般用户。关联规则的发掘基于所有的交易数据，当不同用户购买了不同商品后，根据发掘出来的关联规则，系统提供不同的相关产品，实现广粒度级的个性化服务。

由用户的交易信息，每个订单编号可唯一标识每一次交易事件，所有的订单号构成了交易集合 D，而每次交易的商品名称构成了项目集中的每一个 i，依照 Apriori 算法产生所有的大数据项集，其中每一个大数据项集都是互相有关联的商品。

在所有的大数据项集上产生关联规则，为了使算法提供的个性化信息更加准确，关联规则只产生到后项只有一项的所有规则，这样一方面减少了计算时间，另一方面使得前提尽可能多，结论尽可能少，既可使结论更加准确，又避免了相关信息提供的太多而出现“信息过多，知识过少的现象”。

4.2 用户群级个性化服务

由 Apriori 算法和 Rule-Generation 算法实现了针对一般用户的广粒度级的个性化服务。但数据挖掘算法的一个主要目标就是要降低计算量，提高工作效率，提高服务质量。从硬件上来改善是一个努力方向，改善算法或采用并行计算是另一个方向，而针对具体的应用领域合理地缩小知识发现的数据量也将有效改善数据挖掘算法的工作效率并提供更恰当数据来建立挖掘模式，提供更准确合理的服务。

针对 PeopleFirst 的应用领域，可按照注册用户在网站的个人注册信息性别、年龄、收入、范围、健康需求将会员划分成不同的用户群。而这些离散化(即将年龄离散化为0～1岁，1～5岁，5～12岁，12～20岁，20～30岁，30～50岁，50岁以上等等)的个人信息构成了个人信息空间。

定义1(个人信息空间) 四组个人个性化信息组成了一个四维超立方体空间(若有 n 个指标，则组成 n 维空间)，每一维均由有限的几个离散点构成，这样，由四组个性化信息组成的四维空间所含的元素是有限的，即各维元素个数的连乘积。定义该四维空间为个人信息空间。

个人信息空间的有限点构成了有限的用户群。

针对个人信息空间的有限点来选择数据，对于每一个点选择出来的数据都是确切反应该用户群的交易数据。基于这些更能反映该用户群且数据量更小的数据，采用与4.1节相同的方法来进行数据挖掘。由用户群交易信息挖掘出来的相关商品将符合该用户群的消费水平，例如对其他个性化信息一样，收入较高的人群，关联模式将提供价格稍高且商品质量较好的同一类商品。

这样将有效促进每一用户群数据挖掘算法的计算效率，而且挖掘出来的关联模式将更准确地反映该用户群的相关商品信息。根据不同的用户群建立不同的挖掘模式索引，这样当输入某一个特定用户群时，相应的索引将迅速挖掘相应中粒度级个性化信息。

由上面的分析可以看出，个人信息空间中点的多少与数据计算量是一种彼此消长的关系。即个人信息空间中的点越多，挖掘计算量越小，同时由于分类太多，类的维护开销增大，存储空间增大。寻找分类的多少与计算量大小的最佳平衡点是一个新的研究课题(在此不作详细讨论)，其输入是网站交易信息量的大小，平衡点是多方面因素综合后的一个优化问题。

4.3 ECPI Miner 的设计

ECPI Miner 是一个电子商务个性化信息挖掘工具。它将关联分析的算法集成于其中，输入是用户的个人信息(一般用户或有个人注册信息的某一用户群)和此次购买商品，输出是根据挖掘算法计算出来的相关商品。如图2。

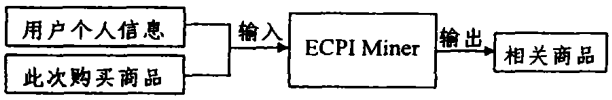

图2 ECPI Miner 工作方式

结论 利用数据挖掘的方法实现了两种不同级别的粗粒度个性化服务。即针对有个人注册信息的会员，可提供基于用户群粒度级的个性化服务。针对一般的非注册用户提供广粒度级的个性化服务，即根据所有的交易信息来计算相关商品的提供模式。

参考文献

1 Li Changqing, Tang Shiwei, Li Hongyan, Zhao Wenbing. Improving on the Personalized Algorithm and Research on the Pertinence of Influence Index in E-Business. In: Proc. of the 9th National Conf. on Multimedia Technology NCMT'2000. Oct. 2000. 285～288

2 Tian Shengfeng, Huang Houkuan. Artificial Intelligence and Knowledge Engineering. China Railway Press, 1999. 8

3 Wang Shan, et al. Data Warehouse Technology and OLAP. Science Presss, 1998. 6

4 Hu Kan, Xia Shaowei. Large Data Warehouse-based Data Mining: a Survery. Journal of Software, 1998, (1): 53～63

5 Mannila H, Tovivonen H, Inkeri Verkamo A. Efficient Algorithm for Discovering Association Rules. In: Proc. of AAAI Workshop on Knowledge Discovery in Database. July 1994. 181～192

6 Srikant R, Agrawal R. Mining Generalized Association Rules. In: Proc. of the 21th Intl. Conf. on Very Large Database. Sept. 1995. 407～419

7 Toivonen H, Klemettinen M, Ronkaine P, et al. Pruning and Grouping Discovered Association Rules. Mlnet Workshop on Statistics, Machine Learning and Discovery in Database. Heraklion, Crete, Greece, April 1995

8 Agrawal R, Srikant R. Fast Algorithm for Mining Association Rules. In: Proc. of the 21th Intl. Conf. on Very Large Database. Santiago, Chile, Sept. 1994

(上接第35页)

8 Fink J, Kobsa A, Nill A. User-oriented Adaptivity and Adaptability in the AVANTI Project. In Designing for the Web: Empirical Studies. 1996

9 Perkowitz M, Etzioni O. Adaptive Web Sites: Automatically Synthesizing Web Pages. In: Proc. of AAAI98. 1998

10 Perkowitz M, Etzioni O. Adaptive web sites: an AI challenge. In: Proc. 15th Int. Joint Conf. AI. 1997a

11 Perkowitz M, Etzioni O. Adaptive web sites: Automatically learning from user access patterns. In: Proc. of the Sixth Int. WWW conference. 5. 1997b

12 Rabiner L R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 1989, 77(2): 257～286

13 Luotonen A. The common log file format. http://www.w3.org/pub/WWW/, 1995

14 Zaiane O R. Resource and knowledge discovery from the internet and multimedia repositories, doctoral dissertation, university of Simon Fraser, 1999

15 Mladenic D. Machine Learning on non-homogeneous, distributed text data, doctoral dissertation, university of Ljubljana, 1998

16 Chen M S, Park J S, Yu P S. Data mining for path traversal patterns in a web environment. In ICDCS, 1996. 385～392

17 Brin S, Page L. The anatomy of a large-scale hypertextual web search engine. In: 7th Int. Conf. WWW, Brisbane, Australia, April 1998

18 Chakrabarti S, et al. Experiments in topic distillation. In: ACM SIGIR Workshop on Hypertext Information Retrieval on the Web, Melbourne, Australia, 1998

19 Wexelblat A, Maes P. Footprints: History-rich web browsing. In: Proc. Conf. Computer-Assisted Information Retrieval (RIAO), 1997. 75～84

20 Spiliopoulou M, Faulstich L C. WUM: A Web Utilization Miner. 1998

21 Spiliopoulou M. The laborious way from data mining to web mining. Int. Journal of Comp. Sys., Sci. & Eng., Special Issue on "Semantics of the Web". Mar. 1999

22 Craven M, et al. Learning to Extract Symbolic Knowledge from the World Wide Web. In: Proc. of the 15th National Conf. on Artificial Intelligence (AAAI-98). 1998