住宅与房地产电子政务中数据挖掘的应用研究

吉同路[1],柏永飞[2],王立松[2]
(1.江苏省建设厅,江苏 南京 210013;
2.南京航空航天大学 信息科学与技术学院,江苏 南京 210016)

摘 要:住宅与房地产电子政务系统中存在大量的数据,虽然满足了业务系统的应用需求,但如何使这些数据实现更有效的共享,并从中抽取出更多有意义的数据和知识是具有行业特色的电子政务系统建设的一个关键内容。文中主要针对住宅与房地产电子政务平台的规划和需求,根据住宅与房地产行业特点,首先建立住宅与房地产电子政务系统的数据仓库,然后设计相关数据挖掘算法进行数据挖掘,得出了较有意义的挖掘结果,为决策者提供了具有辅助决策意义的数据及其分析结果。

关键词:电子政务;数据仓库;数据挖掘
中图分类号:TP39 **文献标识码**:A **文章编号**:1673-629X(2010)01-0239-04

Study and Application of Data Mining in E-government of House and Real Estate Industry

JI Tong-lu[1],BAI Yong-fei[2],WANG Li-song[2]
(1. Construction Bureau of Jiangsu Province, Nanjing 210013, China;
2. School of Information Science & Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract: There are large amounts of data in housing and real estate E-government platform, and it can meet the application requirements of business systems, but it is also a critical content for e-government system on how to achieve an effective data-sharing and extract of more meaningful data and knowledge. Based on E-government platform's designing and demands and the characteristics of the housing and real estate industry, discussed how to apply the data warehouse and data mining techniques to help government dig useful data, and use them to make intelligently analyze and assist decision-making.

Key words: e-government; data warehouse; data mining

0 引 言

研究表明,大约80%以上的重要信息资源掌握在政府手中,如何利用新技术高效、准确地从政府网站上提取数据实现信息有效共享、提高政府决策的科学性和规范性,使数据转变为知识和财富,使网站成为服务公众、实现资源共享、提高效能的电子政府,这是"电子政务"建设的一个核心和关键问题[1]。数据挖掘技术正是为满足这种需要而产生的一种综合技术,它包含了统计学、机器学习、人工智能、数据库、知识获取、模式识别、分布式多媒体环境的智能代理等。在电子政务中合理使用数据挖掘技术能有效地对政府部门丰富的海量"数据资产"进行开采并加以提炼,使之成为有用的知识,从而对政府的决策起到指导、预测作用。

数据仓库是数据挖掘和知识发现的基础,是对原始数据库进行数据清理、筛选后形成的,在深度和广度上都比原始数据更好[2]。电子政务数据仓库是数据挖掘系统的数据准备阶段,是在相当长的时间段内积累了政府在不同时期的各类文档、统计报表等数据资料,在纵向与横向上都为数据挖掘提供了更为广阔的活动空间,为高效率进行数据挖掘算法打下了良好的基础。

所谓数据挖掘,就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中,提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程[3~7]。

收稿日期:2009-05-08;**修回日期**:2009-08-30
基金项目:国家建设部《城市低收入家庭住房保障信息基础数据标准》和《城市低收入家庭住房保障信息技术规范》(建标[2009]88号)资助项目
作者简介:吉同路(1966-),男,硕士,高级工程师,研究方向为电子政务、数据挖掘、软件工程。

数据挖掘常用的技术有关联规则、决策树、粗糙集、神经网络、遗传算法及各种算法的融合等。简言之，数据挖掘技术的思想是从数据中抽取有价值的信息，以帮助决策者寻找数据间潜在的关联，发现被忽略的要素，这些信息有可能对预测趋势和决策行为十分有用。从更广义的角度来讲，数据挖掘就是在一些事实或观察数据的集合中寻找模式的决策支持过程。其目的就是有效地从海量数据中提取出需要的答案，实现"数据—>信息—>知识—>价值"的转变过程。数据挖掘最吸引人的地方是它能建立预测模型而不是回顾型的模型。

1 基于数据仓库的电子政务数据挖掘

数据库管理系统以数据仓库为核心，它对数据库中的事物及数据进行集成、转换和综合，将数据重新组合成面向分析的全局数据视图。它的主体由关系型数据及多维数据构成，将数据库与模型库、方法库有机地结合在一起，以对象的方式进行储存。系统从信息管理系统数据库等内部、本地外部及远程外部数据源中采集到有用数据之后，根据所定义的元数据对数据按主题进行转换及清理，然后放入数据仓库。

数据挖掘与数据仓库联系非常密切，数据仓库为数据挖掘和知识发现提供了源数据，与此同时数据仓库也是数据挖掘的对象。基于电子政务数据仓库的数据挖掘可以发现有关政务知识，其过程是从电子政务数据仓库中选择数据，通过数据接口转换后，在数据挖掘系统管理器里进行挖掘处理。其结构如图1所示。

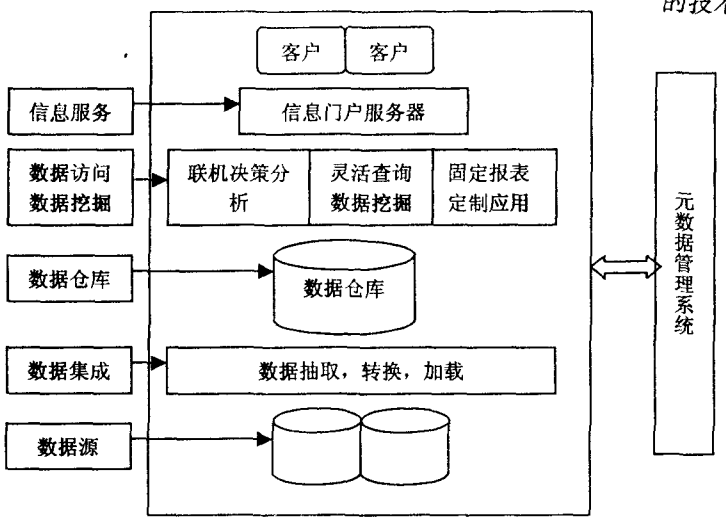

图1 基于数据仓库的数据挖掘模型

2 电子政务中的数据挖掘技术

数据挖掘主要侧重对数据进行分类、聚类、关联、预测分析。数据挖掘在电子政务中的应用就是把分类、聚类、关联、预测这四类分析技术折射到政府部门，使政府部门的内部信息与外部信息进行有效的整合，从而使政府部门更好地、更有效地将信息发布给最希望得到它们的公众，使得政府部门更好地服务于公众。

如果将数据挖掘技术引入到电子政务系统中，就需要在Web服务器上构建一个数据库系统，用来有针对性地记录政府办公人员和民众浏览和操作的路径。该系统包含多个原始的静态数据库，对于系统管理员给出的一个特定的挖掘任务，需要从中抽取进一层的关联数据库，关联数据库及其操作置于后台数据库系统中。其结构如图2所示。

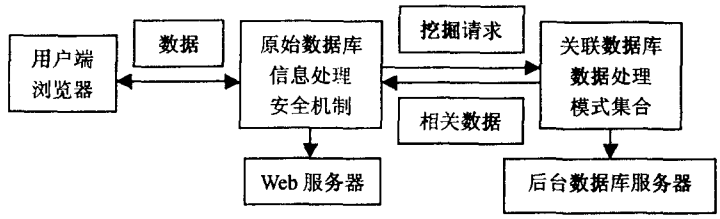

图2 数据挖掘功能和流程

对于以上功能和流程，数据挖掘主要分为以下三个步骤：

(1)数据的提取与净化。

这一过程是制定挖掘任务的过程。向Web服务器提出请求，从已有的数据库中提取相关数据，可用数据库查询语言SQL实现。因为制定挖掘任务时，不同的表达方式可能会造成数据挖掘算法对数据含义理解的不确定性，所以搜集完相关数据后，还需要清除无用数据，对带噪音的数据进行净化。该过程可用统计学的技术检测异常值，进行平滑处理以及估计噪音参数。

(2)数据挖掘算法。

针对不同问题和不同解决方案，上述过程形成了多个由数据挖掘过程中使用到的信息组成的定制数据库，针对这些数据库有很多的数据挖掘算法。几乎所有的数据挖掘算法都要事先提出一些标准来度量产生的模式，通常利用诸如置信度、感兴趣度等的统计属性作为对产生模式的评估标准，从而更进一步决定哪些模式可以保留，哪些模式需要丢弃，更有效地找出潜在的有兴趣的模式。

(3)精化数据和使用结果。

对于运行数据挖掘算法生成的过程，可以循环进行新一轮的数据挖掘过程，同时与分析者以及领域专家进行沟通。反复精化、筛选得到认可后，即成为各种形式的知识，其集合就构成了知识库，可以被用户使用。在此模型中就是将有用信息回

馈给系统,从而帮助他们准确迅速地决策。

3 数据仓库与数据挖掘在住宅与房地产电子政务中的应用

系统使用面向服务(SOA)技术,对全省各级房地产市场交易备案信息系统的日交易数据进行自动采集。在此基础上,通过数据聚集、筛选加工,形成能够满足数据挖掘业务要求的数据仓库。数据仓库建立完成后,根据各个层面对房地产市场报告的要求,结合 BI 技术实现自动分析、智能报告。

3.1 数据采集

根据房产管理信息系统[8]和全省房地产市场交易备案信息采集规范,以及房地产行业行政管理部门有关规定以及国家住宅与房地产政策制定,需要采集的信息包括政府、企业、客户三方面。其中主要有房地产开发企业信息,物业企业信息,评估企业信息,房地产市场交易信息,客户信息等。系统的信息采集步骤分为如下 4 步:

(1)江苏省住房电子政务平台提供 Web Service 接口,定义标准化格式的 XML 文件。

(2)各地房地产市场交易信息备案系统根据 Web Service 接口的具体技术要求,自行生成符合该标准的文件,通过 Web Service 接口来自动上传文件。

(3)江苏省住房电子政务平台对上传的数据文件进行自动分析,将符合要求的数据进行入库处理,返回处理结果。

(4)上传数据单位从江苏省电子政务平台上获取数据上传的情况与质量报告,决定是否进行重传。

3.2 数据自动分析

数据自动分析功能模块包括数据仓库的建立以及数据挖掘模型的建立两个主要部分:

1)建立数据仓库。

系统对数据仓库中数据结构的设计则是在现有业务系统数据结构基础上对数据的名称、类型、描述及关联等进行了重新定义,主要包括统一数据类型、调整数据长度和增加时间属性。

根据数据仓库的要求,对从不同数据源加载来的数据,统一其数据类型并调整其数据长度,以确保数据仓库数据的一致性。

对时间属性的设计,则是为数据仓库中的每一个表设置 2 个日期类型的字段"数据开始日期"和"数据结束日期",由此描述数据所属的时间段。

数据仓库元数据包括对整个数据仓库环境的描述,它在数据仓库的设计、运行中起着极其重要的作用。它描述了数据仓库中的各个对象,遍及数据仓库的所有方面,是整个数据仓库的核心。在系统中,数据仓库的元数据描述了数据仓库环境的数据字典和数据处理规则。数据仓库的设计和数据清理转换均根据元数据,从而通过元数据记录了一份关于数据仓库环境的完整的资料文档。

系统中元数据用关系表描述,元数据的数据库结构包括:

(1)数据源表:保存数据源的有关信息,包括数据源类型、数据库服务器名、数据库名、用户名、用户口令。

(2)数据仓库信息表:保存目的数据仓库的信息。

(3)任务表:保存抽取任务描述信息。

(4)数据映射表:保存抽取表映射和抽取表的字段级映射信息。

(5)数据清洁表:保存有关数据清洁的描述。

(6)数据检验表:保存有关数据相关性检验的描述。

(7)数据综合表:保存有关数据综合的描述。

(8) 数据优化表:保存有关数据优化的描述。

2)数据挖掘的应用。

在数据挖掘领域数据挖掘功能发现的模式类别主要有关联规则、分类、聚类、概念描述和偏差检测等。数据挖掘的结果就主要体现在这些模式的发现上。在电子政务中数据挖掘的常用方法主要有决策树方法、统计的方法、归纳法、神经网络方法、遗传算法、粗糙集方法、人工智能、模糊集方法等。

系统根据住宅与房地产行政管理部门以及客户所关心的问题制定了一系列挖掘方法,主要应用了分类、聚类、概念描述等关联规则。在此主要讨论关联规则挖掘方法。

3.2.1 单维关联规则算法和频繁项集

关联规则挖掘过程实际上就是一个从数据仓库中寻找频繁项集的过程,找出了所有频繁项集也就完成了挖掘的主要任务。系统采用 Apriori 算法产生频繁项集。

该算法分两步:

第一步:连接步,由 L_{k-1} 连接自身产生候选 $K-$项集合。该候选项集的集合记作 C_k。设 l_1 和 l_2 是 L_{k-1} 中的项集。记号 $l_i[j]$ 表示 l_i 的第 j 项,为方便计算,假设事务或项集中的项按字典次序排序。执行 $L_{k-1} \triangleright\triangleleft L_k$,如果 l_1 和 l_2 中的前 $K-2$ 项相同,则 l_1 的第 $K-1$ 项必须小于 l_2 中的 $K-1$ 项,这样保证不会产生重复。

第二步:修剪步,从产生的候选项集 C_k 中去除非频繁项集。据 Apriori 性质,任何非频繁的$(K-1)-$项

集都不可能是频繁 $K-$ 项集。如果候选 $K-$ 项集的($K-1$)- 项集不在 L_{k-1} 中,则此候选不是频繁项集,可以从 C_k 中删除它。然后,扫描数据库,得到 C_k 中每个候选项集的计数,把能够满足最小支持度的候选集选取出来,这就是需要查找的最大频繁集。

算法描述为:

```
procedure apriori L_{k-1} - gen(L_{k-1}:frequent(k - 1) - itemsets;min
- sup:mini_ sum
  support threshold)
  for each itemset l_1 ∈ L_{k-1}
    for each itemset l_2 ∈ L_{k-1}
      if(l_1[1] = l_2[1]) ∧ (l_1[2] = l_2[2]) ∧ ⋯ ∧ (l_1[k-2]
= l_2[k-2]) ∧
      (l_1[k-1] = l_2[k-1])then{
          C = l_1 ▷◁ l_2 //join step,generate candidates
          If has_ infrequent_ subset(C,L_{k-1})then
              delete C;
          else add C to C_k;
          }
    return C_k;
procedure has_ infrequent_ subset(C:candidate K - itemset;
  L_{k-1}:frequent(k - 1) - itemsets)//use prior knowledge
    for each(k - 1) - subsets of C
      if s ∉ L_{k-1} then
          return true;
return false;
```

把以上算法应用到下面的算法中则产生交易数据库频繁项集。

算法:Aprioir 使用根据候选生成的逐层迭代找出频繁项集。

输入:事物数据库 D;最小支持度阈值 min - sup。

输出:D 中的频繁项集 L。

```
L_1 = find_ frequent_ 1 - itemsets(D);
for (k = 2;L_{k-1} ≠ ∅;k ++) {
      C_k = apriori_ gen(L_{k-1},min_ sup);
  for each transaction t ∈ D {//scan D for counts
      C_t = subset(C_k,t);//get the subsets of t that are
candidates for
      each candidate c ∈ C_t
          c.count ++;
    }
    L_k = {c ∈ C_k | c.count ≥ min_ sup}
}
return L = U_k L_k;
```

3.2.2 由频繁项集产生关联规则

由 Apriori 算法找出数据库中的频繁项集,根据频繁项集产生强关联规则即满足最小支持度和最小置信度的规则。

支持度:$\text{support}(A \Rightarrow B) = P(A \cup B)$

置信度:

$$\text{confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{support\_ count}(A \cup B)}{\text{support\_ count}(A)}$$

其中,support_ count($A \cup B$) 是包含项集 $A \cup B$ 交易个数,support_ count(A) 指包含项集 A 的交易个数。可信度可由业内专家给出。根据该式,关联规则可以如下产生:

(1) 对于每个频繁项集 l,产生 l 的所有非空子集。

(2) 对于 l 的每个非空子集 s,如果 $\frac{\text{support\_ count}(l)}{\text{support\_ count}(s)} \geq \text{min\_ conf}$ 则输出规则"$s \Rightarrow (l-s)$"

其中,min - conf 是最小置信度阈值。

3.2.3 结果分析

利用以上算法则可以得到很多有用的规则如表 1 所示:

表 1 一些有用的规则

规则	支持度(%)	可信度(%)
高档别墅⇒临近山水	20.3	40.2
高档商品办公楼⇒(临近市中心,经济开发区)	30.2	40.5
房地产投资增速⇒政策支持	8.5	10.6
房地产投资下降⇒需求下降	10.2	20.4
地理位置无关型客户⇒重视物业管理	9.2	35
地理位置不重要型客户⇒商务活动比较频繁	4.5	42.1
地理位置参考型客户⇒较关注户型	4.3	15.7
地理位置重要型⇒希望社区规模较小	4.7	17.5

可信度反映了关联规则前提成立的条件下结果成立的概率,支持度反映了关联是否是普遍存在的规律。例如对于高档商品办公楼建在市中心或新的经济技术开发区的可信度为 40.5% 且数据的支持率为 30.2%。运用关联分析的目的是寻找数据库中值的相关性,其他被发掘的关联也可以通过类似的比较,进行深一步的挖掘。利用数据挖掘得到的有用信息可以帮助政府提高决策能力,对宏观经济形式及市场发展趋势做出更为精确的判断。

4 结束语

发展电子政务,利用电子政务综合数据库中存储的大量数据,通过数据仓库、数据挖掘技术,建立正确的决策体系和决策支持模型,为各级政府的决策提供科学的依据,从而提高各项政策制定的科学性和合理

(下转第 246 页)

对其进行快速转换，得到相应的 CNMARC(DC)字段的域值，从而提高资源的共享效率。

4.5 资源查询

数据查询是 DRMS 系统前端用户界面的主要功能之一，系统提供交互式查询工具，可进行多条件组合和模糊查询，也可自行设计查询数据表的显示格式，如将“语种”的 CN 字符显示为“中文”，“类型”JN 显示为“过刊”。查询结果可通过多页数据表的形式进行分类，与之关联的数据则可输出显示成文本形式。

4.6 数据分析

数据分析和数据统计是图书馆日常管理的有效方法，也是用户对其所关注数据进行处理的常用方法之一。对 DRMS 的实现方法而言，如果直接将结果输出，一般用户难于理解其真实含义，使得结果分析较为困难。因此，在对检索结果进行处理的同时，还可以对其进行数理统计分析，如统计和预估出读者对于何种文献在特定时间内的需求量，总结不同文献在特定关注条件下所表现出的变化规律和参数特征，为图书馆的文献订阅和服务模式的改进提供直观依据。

5 结束语

数字资源管理系统 DRMS 灵活、方便、实用，为图书馆学应用于信息检索构建了良好的数据分析和数据应用平台。DRMS 系统的数字资源参数数据库、数据库前端用户界面和数据应用接口函数库的设计，方便了不同层面的应用需求面向对象的组件化软件设计，便于数据管理扩容和功能扩充。由于系统不需要商业软件的支持，对硬件环境没有特别的要求，因此，具有较强的灵活性和可移植性，既可以单独运行，也可以作为软件包组装到其它软件系统上使用，取得了良好的应用效果。

参考文献：

[1] 张宇娥.数字图书馆建设中数字资源整合研究[J].电子科技大学学报，2002，31(s)：42－44.

[2] William H，Sameer A，Rodney L L，et al. SPIRS：A Web－based image retrieval system for large biomedical databases [J]. International Journal of Medical Informatics，2009，78(4)：13－24.

[3] Wendy O，Daljit K，Kathy C，et al. A cross－platform solution for bibliographic record manipulation in digital libraries [C]//Proceedings of the Sixth IASTED International Conference on Communications，Internet，and Info. Technology. July 2－4，2007. Banff，AB，Canada：[s. n.]，2007：193－198.

[4] 徐汝兴.图书馆跨平台信息检索系统初探[J].上海交通大学学报，2003，37(s1)：191－194.

[5] 姜爱蓉，黄美君，窦天芳.数字资源整合与信息门户建设[J].现代图书情报技术，2006(11)：2－6.

[6] 段智敏，王如龙，孙美青.基于一卡通的数字化校园资源整合研究与实现[J].计算机工程与科学，2008，30(1)：8－11.

[7] Martin W，Alois K. A cross platform development workflow for C/C++ applications[C]//The 3rd International Conference on Software Engineering Advances，Sliema，Malta：[s. n.]，2008：224－229.

[8] Blanchette J，Mark S. C++ GUI Programming with Qt 4 [M]. 2nd Ed. [s. l.]：Prentice Hall，2008.

[9] Powers L，Snell M. Visual Studio 2005 技术大全[M].刘彦博，肖 鹏，贾 菡，译.北京：人民邮电出版社，2008.

[10] 王 晟. C++ Builder 数据库开发经典案例解析[M].北京：清华大学出版社，2005.

[11] Hollingworth J，Swart B，Cashman M，et al. Borland C++ Builder 6 Developer's Guide[M]. USA：Sams，2002.

[12] 罗庭芝.从编目角度探讨图书馆数字资源的建设[J].当代图书馆，2008(1)：48－51.

(上接第 242 页)

性，以达到提高政府办公效率、促进经济发展的目的。而数据仓库、数据挖掘正是实现政府决策支持的核心技术。以数据仓库、数据挖掘为依托的政府决策支持系统，将发挥重要的作用。

参考文献：

[1] 吉同路.政府资源计划(GRP)初探[J].哈尔滨工业大学学报，2003，35(s)：132－136.

[2] 王 珊.数据仓库技术与联机分析处理[M].北京：科学出版社，1998.

[3] Inmon W H，Hackathorn R D. Using the Data Warehouse [M]. New York：A Wiley Q－ED Publication，1994.

[4] Han J W，Micheline K. Data Mining Concepts a Technique [M]. Beijing：High Education Press，2001.

[5] Biswas G，Weinberg J B，Fisher D H. Iterate：A Conceptual Clustering Algorithm for Data Mining[J]. IEEE Transactions on Systems，Man，and Cybernetics (Part C)，1998，28(2)：100－111.

[6] Chen M S，Han J，Yu P S. Data Mining：An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering，1996，8(6)：866－883.

[7] Glymour C，Madigan D，Pregibon D，et al. Statistical Themes and Lessons for Data Mining[J]. Data Mining and Knowledge Discovery，1997，1(1)：11－28.

[8] 蒋海琴，阎国年，蒋文明，等.房产管理信息系统[M].北京：科学出版社，2007.