Computer Science and Application 计算机科学与应用, 2018, 8(9), 1317-1325
Published Online September 2018 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2018.89142

Hans汉斯

Research of Family Data Fusion Pattern Based on Uncertainty Reasoning

Yanfeng Jin[1,2], Huifeng Zhang[2*], Wei Jin[2], Yu Liu[2], Xueping Wang[2]

[1]School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing
[2]Shijiazhuang Posts and Telecommunications Technical College, Shijiazhuang Hebei
Email: yanfengjin@163.com, *114560731@qq.com

Received: Aug. 18th, 2018; accepted: Aug. 30th, 2018; published: Sep. 6th, 2018

Abstract

Existing data mining models are often unable to associate individuals with family relationships in the same database. The interest inference algorithm, industry inference algorithm and production inference algorithm are proposed based on certainty factor. Personalized Database and magazine subscription database were fused effectively, and generate new standardized database was generated with key data attributes. An identification rule of family structure was developed as the basis of age, gender, surname and other property, and the integration of family structure data was achieved. Experiments and analysis demonstrated the feasibility and effectiveness.

Keywords

Certainty Factor, Support Degree, Data Fusion, Personalized Database

基于不确定性推理的家庭数据融合算法研究

靳艳峰[1,2]，张慧锋[2*]，靳　伟[2]，刘　羽[2]，王雪平[2]

[1]北京邮电大学经济管理学院，北京
[2]石家庄邮电职业技术学院，河北 石家庄
Email: yanfengjin@163.com, *114560731@qq.com

收稿日期：2018年8月18日；录用日期：2018年8月30日；发布日期：2018年9月6日

摘　要

现有的数据融合模式往往无法将同一数据库中具有家庭关系的个体关联起来，使得目标客户的选择存在

*通讯作者。

文章引用: 靳艳峰, 张慧锋, 靳伟, 刘羽, 王雪平. 基于不确定性推理的家庭数据融合算法研究[J]. 计算机科学与应用, 2018, 8(9): 1317-1325. DOI: 10.12677/csa.2018.89142

重复选取的局限性，同时数据属性的缺失也为进一步的决策带来困难。首先从不确定性推理模型出发，设计出用户兴趣、所处行业、产品偏好等的可信度推理算法，利用该算法将个性化数据库、杂志订阅数据库进行有效融合，在完善关键数据属性的基础上，生成新的标准化数据库；以年龄、性别、姓氏等属性为依据，制定家庭结构的识别规则，从而实现家庭结构数据的融合。利用邮政行业数据及自建数据库进行实验和分析，证明了方法的可行性和有效性。

关键词

可信度，支持度，数据融合，个性化数据库

1. 引言

当前，我国零售业经历了飞速的跨越式的发展阶段，社会消费品年均零售总额有了很大突破，百货商场、连锁店、超市、专卖店、购物中心、电子商务等业态相继兴起。尤其是电子商务的崛起，对零售企业提出了严峻的挑战。随着客户数量的增多，客户对服务的质量提出了更高的要求，另外，二维码技术、客户管理系统、销售分析系统、智能终端设备等的出现，使得零售企业很难在积累的超大规模数据中找出有利于企业开展营销活动的信息。大数据时代的来临也给企业带来了诸如查询响应时间、查询质量、决策分析等多方面的问题[1] [2] [3]。事实上，零售企业在自身的发展中累积了很多的交易数据，最重要的就是要从这些交易数据中找到顾客的购买兴趣和偏好。为了探寻大规模数据背后的隐藏知识，数据挖掘技术应运而生。在国外，对数据挖掘技术的研究已经有多年的历史。数据挖掘算法的发展及其应用，为零售行业带来了巨大的变化，尤其是利润的提高，数据挖掘这一智能化的技术正在被许多的科研院研究[4] [5] [6]。

目前对于可信度模型研究相对较多，而专门针对家庭数据融合的研究相对较少。其中较为代表性的钟诚等通过考察社会网络中人际交往的特征，基于本体理论，提出了在语义网中计算信息可信度的一种方法[7]。徐国虎等以国共合作领域本体库为实例，分析领域本体中四种常见语义关系的特点，在此基础上定义了推理规则，并采用 JRL 规则语言形式化了本体推理规则[8]。王梁等提出溯源表达式的转换算法，对元组中多个属性存在不确定性的情况，可以正确计算结果元组的概率[9]。李艳娜等首次将证据理论和本体相结合，提出了基于证据理论的不确定性上下文本体建模方法，并对证据组合规则进行了修改，解决了证据理论在高度证据冲突时的局限性，设计并实现了具有自适应性的不确定上下文推理算法[10]。李慧芳等将可信度方法与本体模型相结合，提出了基于可信度修正的不确定性推理方法，在推理过程中根据证据的可靠情况实时地修正知识的可信度，并利用修正后的知识可信度值来计算结论的综合可信度[11]。综合上述，专门针对家庭结构模式挖掘的研究还有所欠缺，因此本文将在不确定性推理的基础上，对家庭数据融合模式进行实证研究。

2. 可信度推理与算法设计

2.1. 可信度推理

可信度方法是 MYCIN 系统采用的一种不精确的推理模型，在许多实际应用中都是一个合理有效的

推理模式[12]。根据经验对一个事物或现象为真的相信程度称为可信度。每条规则具有一个可信度，每个证据也具有一个可信度。

IF E THEN H ($CF(H, E)$)，其中 $CF(H,E)$ 为该规则的可信度。

$CF(H,E)$ 表示已知证据 E 的情况下对假设 H 为真的支持程度，其取值在[−1, 1]。

$CF(H,E)>0$ 表示结论为真的程度，值越大 H 越真。$CF(H,E)=1$ 表示为真。

$CF(H,E)<0$ 表示结论为假的程度，值越小 H 越假。$CF(H,E)=-1$ 表示为假。

$CF(H,E)=0$ 表示 E 与 H 没有关系。

定义 $CF(H,E)=MB(H,E)-MD(H,E)$。其中 $MB(H,E)$ 称为信任增长度，表示因证据 E 的出现而对假设 H 为真的信任增加程度。$MD(H,E)$ 称为不信任增长度，表示因证据 E 的出现而对假设 H 为真的信任减少的程度。

当 $MB(H,E)>0$ 时，$P(H|E)>P(H)$。当 $MB(H,E)<0$ 时，$P(H|E)<P(H)$。因此得到：

如果 $P(H|E)<P(H)$，则：$MB=0$，$MD=\left[P(H)-P(H|E)\right]/P(H)$，$CF=MB-MD$。

如果 $P(H|E)>P(H)$，则：$MD=0$，$MB=\left[P(H|E)-P(H)\right]/\left[1-P(H)\right]$，$CF=MB-MD$。

2.2. 推理算法设计

对于一般的客户数据库而言，很少能真正找到个体的兴趣及其所从事的行业属性，本文考虑从其他相关联的数据库中，通过一定的推理，用定量的方法找到其兴趣、从事行业以及偏好的产品种类[13] [14] [15]。

定义 1 个性化数据库(Personalized Database)：是指包含了个人基本信息的不同群体的数据库。比如教师库，包含了某个区域范围内的所有教师的基本信息，如姓名、年龄、家庭住址等基本信息。

定义 2 杂志订阅数据库(Magazine Subscriptions Database)：是指包含了某个区域范围内订阅杂志用户的基本信息的数据库。其中的数据属性包括姓名、家庭住址、联系方式、订阅杂志名称等。

定义 3 标准数据库(Standard Database)：是指融合了个性化数据库、杂志订阅库、小区数据库等数据库信息的数据库。该数据库除了包含上述数据库种的信息之外，还包含了通过可信度推理算法得出的新的数据属性，比如兴趣、从事行业、产品偏好等信息。

用户兴趣的挖掘主要依据杂志订阅库中用户所订阅的杂志信息[16]。首先利用霍兰德职业兴趣测量表，建立兴趣数据库，其中包含了 21 种兴趣种类。具体参见表 1。其次建立兴趣关键字库，利用文本挖掘方法将不同兴趣对应的关键字提取出来，并建立相应的关键字库。

算法一：兴趣可信度算法

1) 计算每个杂志对应兴趣的关键字的个数 $L_1, L_2, L_3, \cdots, L_{21}$；

2) 兴趣关键字数为 T_i，a_i 为杂志 M_i 在兴趣 I_i 中关键字出现频率，$a_i=\frac{L_i}{T_i}(i=1,2,\cdots,21)$；

3) 计算杂志 M_j 中所有兴趣的后验概率 $p(I_i|M_j)=a_i\Big/\sum_{i=1}^{21}a_i$。$(i=1,2,\cdots,21; j=1,2,\cdots,n)$；

Table 1. Interest types extracted from Hollander's occupational interest list
表 1. 依据霍兰德职业兴趣表提取的兴趣种类

兴趣种类提取						
运动	汽车	音乐	文学	绘画	书法	舞蹈
旅游	购物	影视	上网	理财	交友	下棋
园艺	数码	美容	保健	饮食	集邮	摄影

4) 计算各个兴趣的先验概率 $p(I_i)=1/21$；

5) 计算信任增长度 MB 和不信任增长度 MD：

如果 $p(I_i|M_j)>p(I_i)$，则 $MB_{ji}=\dfrac{P(I_i|M_j)-P(I_i)}{1-P(I_i)}$，$MD_{ji}=0$，$(i=1,2,\cdots,21;j=1,2,\cdots,n)$。

如果 $p(I_i|M_j)<p(I_i)$，则 $MB_{ji}=0$，$MD_{ji}=\dfrac{P(I_i)-P(I_i|M_j)}{P(I_i)}$，$(i=1,2,\cdots,21;j=1,2,\cdots,n)$：

6) 计算兴趣在杂志中的可信度：$CF_{ji}=MB_{ji}-MD_{ji}$。

用户所从事行业的挖掘主要依据也是杂志订阅库中用户所订阅的杂志信息。首先利用霍兰德职业兴趣测量表，建立行业数据库，其中包含了 25 种行业种类。具体参见表 2。其次建立行业关键字库，利用文本挖掘方法将不同行业对应的关键字提取出来，并建立相应的关键字库。

算法二：行业可信度算法

1) 计算每个杂志对应行业的关键字的个数 $L_1,L_2,L_3,\cdots,L_{25}$；

2) 行业关键字数为 T_i，a_i 为杂志 M_i 在行业 H_i 中关键字出现频率，$a_i=\dfrac{L_i}{T_i}(i=1,2,\cdots,25)$；

3) 计算杂志 M_j 中所有行业的后验概率 $p(H_i|M_j)=a_i\Big/\sum_{i=1}^{25}a_i$，$(i=1,2,\cdots,25;j=1,2,\cdots,n)$。

4) 计算各个行业的先验概率 $p(H_i)=1/25$

5) 计算信任增长度 MB 和不信任增长度 MD：

如果 $p(H_i|M_j)>p(H_i)$，则 $MB_{ji}=\dfrac{P(H_i|M_j)-P(H_i)}{1-P(H_i)}$，$MD_{ji}=0$，$(i=1,2,\cdots,25;j=1,2,\cdots,n)$，

如果 $p(H_i|M_j)<p(H_i)$，则 $MB_{ji}=0$，$MD_{ji}=\dfrac{P(H_i)-P(H_i|M_j)}{P(H_i)}$，$(i=1,2,\cdots,25;j=1,2,\cdots,n)$。

6) 计算行业基于杂志的可信度：$CF_{ji}=MB_{ji}-MD_{ji}$。

同样，用户偏好产品的挖掘主要依据也是杂志订阅库中用户所订阅的杂志信息。首先根据霍兰德职业兴趣测量表，建立杂志中包含广告对应的产品数据库，其中包含了 22 种产品广告种类。其次建立产品的广告关键字库，利用文本挖掘方法将不同产品广告对应的关键字提取出来，并建立相应的关键字库

算法三：产品偏好可信度算法

1) 计算每种杂志对应某种产品广告的关键字的个数 $L_1,L_2,L_3,\cdots,L_{22}$。

2) 广告关键字数量为 T_i，a_i 为杂志 M_i 在广告 A_i 中关键字出现频率，$a_i=\dfrac{L_i}{T_i}(i=1,2,\cdots,22)$；

3) 计算杂志 M_j 中所有兴趣的后验概率 $p(A_i|M_j)=a_i\Big/\sum_{i=1}^{22}a_i$，其中 $(i=1,2,\cdots,22;j=1,2,\cdots,n)$；

Table 2. Industry categories extracted from Hollander occupational interest table
表 2. 依据霍兰德职业兴趣表提取的行业种类

行业关键字提取								
IT 业	通讯	汽车	中介服务	教育/培训	批发/零售	环保	学术/科研	农林牧渔
交通运输	酒店餐饮	办公设备	医药/医疗	娱乐/体育	旅游/餐饮/	电力/水利	加工制造	政府/公共事业
能源/矿产/采掘	金融/保险/投资/	医疗/护理/美容	房产/建筑/建材	广告/会展/公关	媒体/出版/影视/文化	机械机电设备	家居/室内设计/装饰	null

4) 计算各种产品广告的先验概率 $p(A_i)=1/22$。

5) 计算信任增长度 MB 和不信任增长度 MD：

如果 $p(A_i \mid M_j) > p(A_i)$，则 $MB_{ji}=\dfrac{P(A_i \mid M_j)-P(A_i)}{1-P(A_i)}$，$MD_{ji}=0$，$(i=1,2,\cdots,22; j=1,2,\cdots,n)$。

如果 $p(A_i \mid M_j) < p(A_i)$，则 $MB_{ji}=0$，$MD_{ji}=\dfrac{P(A_i)-P(A_i \mid M_j)}{P(A_i)}$，$(i=1,2,\cdots,22; j=1,2,\cdots,n)$；

6) 计算产品广告在杂志中的可信度：$CF_{ji}=MB_{ji}-MD_{ji}$。

3. 家庭融合规则设计

基于上述融合形成的标准数据库，依据不同的数据属性，进行数据的家庭融合及类别的划分。可以按照地址、性别、年龄等属性进行条件判断，家庭结构的类型划分为两人结构、三人结构、四人结构以及多人结构情况。四人结构和多人结构规则的设计原理与二人、三人结构类似，由于计算相对比较复杂，这里不做进一步讨论。

3.1. 两人结构家庭的判断规则

一般的两人结构家庭，首先通过地址进行关联，当两条数据记录的地址相同时，初步判定为一个两人结构家庭。在此基础上进行家庭关系的判断，一般在不考虑权重的情况下按照年龄、性别和姓氏进行判断。具体的规则如表 3。

对于标准数据库中地址相同的两人，如果其中一人的年龄属性依照一定的概率来源于某一类型的个性化数据库，此时需要对权重进行设置，由于个性化数据库的类型较多，本文仅以学生个性化数据库为例，进行规则的设计，具体见表 4。

3.2. 三人结构家庭判断规则设计

排版后为通栏的统一调整到页面的顶端或底端；如果是单栏则应调整到页面的左右四角的位置对于三人家庭结构而言，与两人结构家庭的判断规则类似，首先也要通过地址进行关联，当出现三条数据记

Table 3. Rule of two person family structure judgment
表 3. 两人家庭结构判断规则

if(条件)	weight (权重)	家庭结构
−10 < age1 − age2 < 10，sex1! = sex2		夫，妻
23 < age1 − age2 < 35，sex1 = 男，sex2 = 男		父，子
23 < age1 − age2 < 35，sex1 = 男，sex2 = 女		父，女
23 < age1 − age2 < 35，sex1 = 女，sex2 = 男		母，子
23 < age1 − age2 < 35，sex1 = 女，sex2 = 女		母，女
45 < age1 − age2 < 55，sex1 = 男，sex2 = 男，last name1 = last name2	不考虑	爷爷，孙子
45 < age1 − age2 < 55，sex1 = 男，sex2 = 女，last name1 = last name2		爷爷，孙女
45 < age1 − age2 < 55，sex1 = 男，sex2 = 男，last name1! = last name2		外公，外孙
45 < age1 − age2 < 55，sex1 = 男，sex2 = 女，last name1! = last name2		外公，外孙女
45 < age1 − age2 < 55，sex1 = 女，sex2 = 男		[50%奶奶，孙子] [50%外婆，外孙]
45 < age1-age2 < 55，sex1 = 女，sex2 = 女		[50%奶奶，孙女] [50%外婆，外孙女]

Table 4. Rule of two family structure estimation based on probability dependence
表 4. 概率依赖型两人家庭结构判断规则

if(条件)	weight (权重)	age	家庭结构	
		age2X 结果	条件	家庭结构结论
id age1来源于学生库	0.95	age2X = age1 + 27	sex1 = 男，sex2 = 男	父，子
			sex1 = 女，sex2 = 男	父，女
			sex1 = 男，sex2 = 女	母，子
			sex1 = 女，sex2 = 女	母，女
	0.05	age2X = age1 + 50	sex1 = 男，sex2 = 男，last name1 = last name2	祖父，孙子
			sex1 = 男，sex2 = 男，last name1! = last name2	外祖父，孙子
			sex1 = 女，sex2 = 男，last name1 = last name2	祖父，孙女
			sex1 = 女，sex2 = 男，last name1! = last name2	外祖父，孙女
			sex1 = 男，sex2 = 女	(外)祖母，孙子
			sex1 = 女，sex2 = 女	(外)祖母，孙女

录的地址相同时，可以初步判定为一个三人结构家庭。在此基础上进行家庭关系的判断，一般在不考虑权重的情况下按照年龄、性别和姓氏进行判断。具体的规则如表 5 所示。

对于标准数据库中通过地址关联，得到的具有相同地址的三条数据，如果其中两条数据中的年龄属性依照一定的概率来源于某一类型的个性化数据库，此时需要对权重进行设置，由于个性化数据库的类型较多，本文仅以老年人(夕阳红)个性化数据库为例，进行规则的设计，具体见表 6。

上述规则作为标准数据中家庭融合的依据，可将标准化数据库中的数据按照家庭进行组合，并对家庭结构做出判断，为营销策略的制定提供决策支持。

4. 仿真实验及分析

本文将邮政行业的个性化数据库、杂志订阅数据库作为基础数据，根据可信度推理的原理，利用 2 中设计的算法和 3 中制定的融合规则，运用 SQLsever2003 平台，进行了模拟仿真实验。其中个性化数据库包含学生库和夕阳红库的 5879 条数据，杂志订阅库包含了杂志订阅者的 9813 条数据，同时针对杂志订阅库中的 2856 种杂志，自建了兴趣关键字库、行业关键字库、产品广告关键字库等三个数据库。

依据上述数据库中的数据，根据不同用户订阅的杂志信息，利用算法 1 和 2 得到了不同用户的兴趣和行业的可信度值。根据上述结果很容易可以看出，对于数据挖掘后的文本信息，可以利用不确定系统进行计算，从而得出定量数值，用于决策支持。不同类型的兴趣可信度值见表 7，利用算法 1 计算出不同兴趣的可信度值，并按可信度值的大小进行排序，比如《北京大学教育评论》杂志的订阅者具有文学兴趣的可信度值为 0.8，具有教育兴趣的可信度值为 0.5，具有绘画兴趣的可信度值为 0.1，同理可以计算出 21 种不同兴趣的可信度值。

不同类型行业的可信度值见表 8：利用算法 2 计算出不同行业的可信度值，并按可信度值的大小进行排序。比如《财经界》杂志订阅者所属金融行业的可信度值为 0.89，所属家居行业的可信度值为 0.07，而所属制造行业的可信度值为 0，同理可计算出 25 个行业的可信度值。

在上述可信度推理的基础上，将个性化数据库与杂志订阅数据库进行融合，得到了包含用户兴趣、从事行业、产品偏好等属性的标准数据库，利用 3 中的融合规则，从 15,692 条数据中挖掘得到 862 组两人结构家庭，533 组三人结构家庭，123 组四人结构家庭。具体家庭结构分布结果见图 1。

Table 5. Rule of three person family structure judgment
表 5. 三人家庭结构判断规则

if(条件)	weight (权重)	家庭结构
23 < (age1 + age2)/2 − age3 < 35，−10 < age1 − age2 < 10，min(age1,age2) > 23, max(age1,age2) < 50，sex3 = 男	null(不考虑)	男户主，妻子，儿子
23 < (age1 + age2)/2 − age3 < 35，−10 < age1 − age2 < 10，min(age1,age2) > 23, max(age1,age2) < 50，sex3 = 女		男户主，妻子，女儿
sex1 = 男，sex2 = 男，sex3 = 男，23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10		男户主，儿子1，儿子2
23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10，sex1 = 男，sex2！= sex3		男户主，儿子，女儿
23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10，sex1 = 女，sex2 = 男，sex3 = 男		女户主，儿子1，儿子2
23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10，sex1 = 男，sex2 = 女，sex3 = 女		男户主，女儿1，女儿2
23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10，sex1 = 女，sex2！= sex3		女户主，儿子，女儿
23 < age1 < 50，23 < age1 − (age2 + age3)/2 < 35，−10 < age2 − age3 < 10，sex1 = 女，sex2 = 女，sex3 = 女		女户主，女儿1，女儿2
23 < (age1 + age2)/2 − age3 < 35，min(age1,age2) > 50，sex3 = 男		父，母，男户主
23 < (age1 + age2)/2 − age3 < 35，min(age1,age2) > 50，sex3 = 女		父，母，女户主
(age1 + age2)/2 − age3 > 50，min(age1,age2) > 50，age3 < 18，last name1 = last name3，sex3 = 男		爷爷，奶奶，孙子
(age1 + age2)/2 − age3 > 50，min(age1,age2) > 50，age3 < 18，last name1 = last name3，sex3 = 女		爷爷，奶奶，孙女
(age1 + age2)/2 − age3 > 50，min(age1,age2) > 50，age3 < 18，last name1！= last name3，sex3 = 男		外公，外婆，外孙
(age1 + age2)/2 − age3 > 50，min(age1,age2) > 50，age3 < 18，last name1！= last name3，sex3 = 女		外公，外婆，外孙女

Table 6. Rule of three family structure estimation based on probability dependence
表 6. 概率依赖型三人家庭结构判断规则

if(条件)	weight (权重)	age	家庭结构	
		age3X 结果	条件	家庭结构结论
id age1，id age2都来源于夕阳红库	0.65	age3 = (age1 + age2)/2 − 23	sex3 = 男	父，母，男户主
			sex3 = 女	父，母，女户主
	0.35	age3 = (age1 + age2)/2 − 50	sex3 = 男，last name1 = last name3	爷爷，奶奶，孙子
			sex3 = 女，last name1 = last name3	爷爷，奶奶，孙女
			sex3 = 男，last name1！= last name3	外公，外婆，外孙
			sex3 = 女，last name1！= last name3	外公，外婆，外孙女

从上述结果可以看出，对于属性不完全或者不够丰富的数据库数据，很难为企业的营销决策提供有价值的参考，利用可信度推理算法将个性化数据库属性进行完善的同时，根据家庭融合规则，可以得到不同结构家庭单元信息，将其应用到企业营销决策中，不但节约了营销成本，增强针对性，对企业发展

Table 7. Part of interest credibility value
表 7. 部分兴趣可信度值

信息来源	兴趣						
	兴趣 1		兴趣 2		兴趣 3		兴趣 n
	兴趣名称	可信度值	兴趣名称	可信度值	兴趣名称	可信度值	……
北京大学教育评论	文学	0.8	教育	0.5	绘画	0.1	……
财经界	理财	0.6	经济	0.6	文学	0.2	……
长三角	经济	0.9	理财	0.4	文学	0.1	……

Table 8. Part of the industry credit value
表 8. 部分行业可信度值

信息来源	行业							
	行业 1		行业 2		行业 3		行业 n	
	行业名称	可信度值	行业名称	可信度值	行业名称	可信度值	行业名称	可信度值
北京大学教育评论	教育/培训	0.97	医药/医疗器械	0.15	能源/矿产/采掘/冶炼	0.11	……	……
财经界	金融/保险/投资/基金/	0.89	家居/室内设计/装饰装潢/	0.07	加工制造	0	……	……
长三角	教育/培训	1	房产/建筑/建材/工程	0	能源/矿产/采掘/冶炼	0	……	……
船艇	教育/培训	0.86	汽车	0.17	加工制造	0.1	……	……
大众数码	通讯	0.92	娱乐/体育/休闲	0.23	医疗/护理/美容/保健	0	……	……
当代电视	通讯	0.88	娱乐/体育/休闲	0.21	广告/会展/公关	0.18	……	……
当代人	批发/零售	0.79	家居/室内设计/装饰装潢/		广告/会展/公关	0.02	……	……
当代世界	加工制造	0.69	IT 业	0	通讯	0	……	……

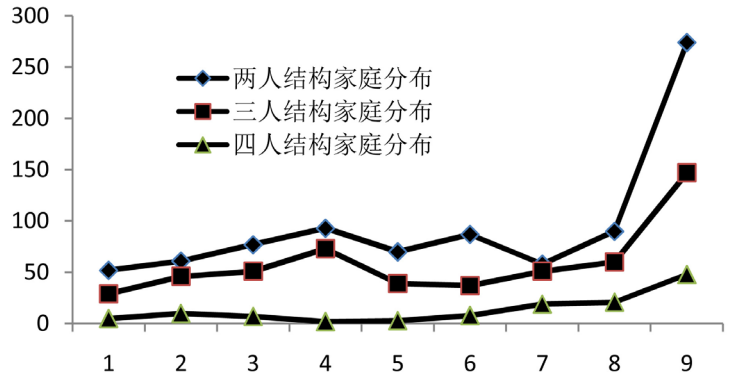

Figure 1. Family structure map
图 1. 家庭结构分布图

意义重大。

5. 结论

针对多数企业客户数据不完善、关键数据属性缺失等情况，本文从两个方面进行了深入研究。一是

数据属性的完善和数据库的标准化。首先通过可信度推理的方式增加数据的属性，对分散的数据库进行整合。文中主要通过对邮政行业的个性化数据库、杂志订阅数据库的挖掘与整合，将缺失的兴趣、行业等关键属性的数据在相应的数据库中进行了补充和完善，其中属性完善主要是根据不确定性推理，计算出相关属性的可信度值。二是家庭结构的识别。基于标准化的数据库，利用地址、年龄、性别、姓氏等属性，设计出了家庭结构的识别规则。首先是按照家庭成员的数量，对家庭结构进行了合理分类，整体划分为两人家庭、三人家庭、四人家庭以及多人家庭，然后根据地址属性来确定家庭结构，根据年龄、性别和姓氏等属性来进一步判断家庭关系，通过实验得出了较为理想的结果。与其他算法相比，文中算法更加有效，实验结果更加符合实际。

基金项目

1) 2018 河北省高校科技计划青年基金项目(编号 QN2018304)；2) 2018 河北省人力资源和社会保障研究课题(编号 JRS-2018-8087)；3) 2018 河北省人力资源和社会保障研究课题(编号 JRS-2018-8088)。

参考文献

[1] 宁彬. 数据挖掘技术及应用研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2005.

[2] 江贺. 聚类若干问题的分析与研究[D]: [硕士学位论文]. 大连: 大连理工大学, 2005.

[3] 贺宗梅. 数据挖掘在网上商店中的应用研究[J]. 科学技术与工程, 2009, 9(10): 2781-2784.

[4] Hahsler, M. (2006) A Model-Based Frequency Constraint for Mining Associations from Transaction Data. *Data Mining and Knowledge Discovery*, **13**, 137-166. https://doi.org/10.1007/s10618-005-0026-2

[5] Cohen, M.D. (2004) Exploiting Response Models: Optimizing Cross-Sell and Up-Sell Opportunities in Banking. *Information Systems*, **29**, 327-341. https://doi.org/10.1016/j.is.2003.08.001

[6] Cartwright, P. (2009) Retail Depositors, Conduct of Business and Sanctioning. *Journal of Financial Regulation and Compliance*, **17**, 302-317.

[7] 钟诚, 赵志峰, 姚高峰. 语义环境中信息可信度计算方法研究[J]. 情报理论与实践, 2012, 35(1): 103-105.

[8] 徐国虎, 许芳, 董慧. 基于语义关系的本体推理规则研究[J]. 中国图书馆学报, 2007, 33(177): 88-92.

[9] 王梁, 周光焱, 王黎维, 等. 不确定关系数据属性级溯源表示与概率计算[J]. 软件学报, 2014, 25(4): 863-879.

[10] 李艳娜, 乔秀全, 李晓峰. 基于证据理论的上下文本体建模以及不确定性推理方法[J]. 电子与信息学报, 2010, 32(8): 1806-1811.

[11] 李慧芳, 张平, 黄鸿, 等. 基于可信度本体模型及不确定性推理的情境感知应用[J]. 北京理工大学学报, 2013, 33(11): 1145-1150.

[12] 岳昆, 刘惟一, 朱运磊, 等. 一种基于概率图模型的不确定性数据世系表示方法[J]. 计算机学报, 2011, 34(10): 1897-1906.

[13] 卫贵武, 黄登仕, 魏宇. 对方案有偏好的不确定语言多属性决策方法[J]. 管理学报, 2007, 4(5): 575-579.

[14] 何亚群, 胡寿松. 不完全信息的多属性粗糖决策分析方法[J]. 系统工程学报, 2004, 19(2): 117-120.

[15] 郑季良, 部平. 决策分析的发展和应用[J]. 云南师范大学学报, 2005, 37(5): 66-71.

[16] 卫贵武. 不确定语言多属性决策的组合方法[J]. 模糊系统与数学, 2008, 22(4): 106-111.

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8801，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：csa@hanspub.org