一种新的重名消解算法在保险领域中的应用研究 *

姚宇峰
(常熟理工学院 计算机科学与工程学院, 江苏 苏州 215500)

摘 要: 研究客户重名消解问题。针对以往重名消解方法如文本聚类的方法需考虑大量无用词汇并需要人工设定阈值以及类别数量,而基于信息抽取的人物相关属性相似度方法对于人物信息的抽取具有依赖性,提出了一种改进的重名消解算法。该算法首先对具有相同标志的客户进行属性匹配,合并匹配成功的标志;然后进行链接分析,对客户合作网的结构进行分析,将具有相同标志并与同一个代理人实体合作的客户归为一个客户实体,并把具有相同合作对的信息加以分析合并;最后通过原子团簇分析法进行聚类分析。仿真实验结果表明,所提改进算法对中文字符串的匹配处理进行了优化,执行效率高,适合于以大量数据为特征的保险领域的重名消解。

关键词: 重名消解; 数据挖掘; 保险领域; 实体
中图分类号: TP391; TP311 **文献标志码**: A **文章编号**: 1001-3695(2012)03-0994-04
doi:10.3969/j.issn.1001-3695.2012.03.053

Application of data mining in customer name disambiguation of insurance field

YAO Yu-feng
(*College of Computer Science & Engineering*, *Changshu Institute of Technology*, *Suzhou Jiangsu* 215500, *China*)

Abstract: This paper researched the solution to customer name disambiguation of the field of insurance. Aiming at the former name disambiguation methods such as text clustering method need to be considered in a lot of useless words, manually set the threshold, and gave he numbers of type, and the method of character-related properties of similarity based on information extraction depends on the character information, proposed a new name disambiguation method. Firstly, applied the same attribute matching, merging the identity of a successful match and then used link analysis, analyzed structural analysis of customers network, the entities had the same identity and classified cooperate with the same policy to a customer entity, merged the same cooperating information. Finally, analyzed cluster analysis cluster. Experiment results show that the proposed method can optimize the chinese text string matching process and have the high implementation efficiency, especially suitable for large amounts of data to the insurance sector is characterized by digestion of the same name.

Key words: name disambiguation; data mining; insurance field; entity

0 引言

近年来,随着养老和医疗等涉及人身保障的各项制度改革的推进,各种核心业务处理系统应运而生,但是在核心业务处理系统数据库中,由于记录信息的局限性,无法找到对应真实客户的唯一主键。那么首先需要面对的问题就是保险业务处理中的客户重名问题。

数据挖掘(datamining)[1,2]是按照既定的业务目标,对大量的数据进行检索,揭示隐含其中的规律并进一步将之模型化的先进有效的方法,它是从大量的、不完全的、有噪声的、模糊的、随机的数据中提取隐含在其中的人们事先又不知道的,但又是潜在的有用信息和知识的过程。重名消解就是根据上下文或篇章信息来区分同一人名表示的不同人物的过程。目前已有的能对重名进行聚类处理的系统主要有 SnakeT、Vivisimo 和 Apex 等,其缺点是仅仅将人名作为普通词汇处理,聚类结果的标签也只是和人名相关的词汇,不能实现对人名的重名结果进行区分。郎君等人[3]提出了一种基于社会网络的人物重名消解方法,Bollegala 等人[4]通过计算向量相似度从而实现最终结果的聚类,文献[5,6]将需要进行重名消解的内容表示成向量空间模型,对重名进行消解。于满泉在文献[3]的基础上进一步通过提取内容中一些相关信息,如名族、性别、学历、工作单位、家庭关系等,创建人物的属性集,在此基础上计算人物的相似度,实现了人物的重名消解。以上方法均考虑了很多无用词汇,并且消解过程对信息的抽取具有很强的依耐性。

本文提出了一种新的重名消解方法,此消解方法分为三个步骤:a)属性匹配,利用属性匹配算法来计算重名客户之间的相似度;b)链接分析,借助于各客户之间的关系来对客户的重名问题进行识别;c)通过原子团簇分析来改进传统的聚类算法。实验结果显示,此方法能对中文字符的匹配处理进行优化,执行效率高,适合于以大量数据为特征的保险领域的重名消解。

1 系统框架和算法原理

1.1 问题描述

人名检索结果的重名消解在保险领域是一个常见的问题,

收稿日期: 2011-08-24; **修回日期**: 2011-09-28 **基金项目**: 常熟理工学院青年教师科研启动基金资助项目(QZ0912)
作者简介:姚宇峰(1981-),男,江苏常熟人,讲师,硕士,主要研究方向为通信运营支撑软件、数据挖掘、图形图像处理等(yaoyufeng3132@163.com).

目前主要采用聚类算法对重名进行消解。可以采用类似于自然语言处理中词义消歧的方法,利用人名的上下文信息来实现。常见的方法将人名检索结果对应的 Snippet 或者网页内容采用向量空间模型表示,或抽取上下文中的关键性短语,采用计算向量相似度的方法来实现最终的检索结果聚类,然后在人物属性集上计算人物的相似度,从而实现人物同一性判别。

保险领域的原始数据信息记录了合同号、投保人姓名、投保人年龄、投保人职业、投保人单位等信息。以下是定义的几个变量:

a)标志。从原始数据所抽取出的合作网络的节点,是客户信息的载体,用字母 r 表示。

b)实体。客户标志所对应的真实的客户,与标志构成一对多的关系,用字母 e 表示。

c)属性。描述作者的属性,如地址、职业等,用字母 a 表示,如 $r_i.a_j$ 表示 r_i 标志 a_j 的属性。

d)$r_i \equiv r_j$。表示标志 r_i 与 r_j 对应同一个实体。

e)关系。两个实体之间的联系称之为关系,关系是一个比较宽泛的概念。

定义　保险领域的重名消解问题。对于给定的客户标志集合 S_r,合作关系集合 $S_c=\{(r_i,r_j)\}$,对于 S_r 中具有某些属性相同、相近或相似的客户标志 r_i、r_j,求集合 $S_e=\{e_i\}$。其中,S_e 表示一个实体集合,e_i 表示与客户标志相对应的实体。

1.2　重名消解步骤

重名消解系统处理过程主要分为三步:a)属性匹配,对相同标志的客户信息的属性进行匹配,并合并匹配成功的标志;b)链接分析,通过对客户合作网的结构分析,发现具有相同标志的客户与同一个代理人实体合作,那么这些具有相同标志的作者标志对应同一个作者实体,并把具有相同合作对的信息加以分析合并;c)原子团簇分析法进行聚类分析。

重名消解流程如图 1 所示。

1.3　属性匹配

对于已经在某保险公司进行投保的客户,其属性有客户姓名、购买产品名称以及客户的信息等。首先对于输入实体的两个属性 $r_i.\text{attr}$ 和 $r_j.\text{attr}$ 进行分割,对两者的第一部分进行子串匹配,如式(1)所示:

$$\text{SubStringMatch}(r_i.\text{Attr.FirstPart},r_j.\text{Attr.FirstPart},R_l) \tag{1}$$

其中:R_l 是一个取值范围为[0,1]的阈值。假设 $r_i.\text{attr.FirstPart}$ 和 $r_j.\text{attr.FirstPart}$ 中较短的字符串为 s、较长的字符串为 l,同时假设 s 在 l 中出现的最长子串为 m,则可以得到

$$\text{SubStringMatch}=\begin{cases}\text{true}, s.\text{LENGTH}\leqslant m.\text{LENGTH}\times R_l\\ \text{false}, s.\text{LENGTH}>m.\text{LENGTH}\times R_l\end{cases} \tag{2}$$

当式(1)返回为 True 时,再通过式(3)进行第二部分匹配:

$$\text{CountMatch}(r_i.\text{Attr},r_j.\text{Attr},R_c) \tag{3}$$

其中:R_c 同样是取值范围为[0,1]的阈值。假设 $r_i.\text{Attr}$ 和 $r_j.\text{Attr}$ 中较短的字符串为 s、较长的为 l,同时假设 s 中的字符在 l 中出现的次数为 n,则可得到

$$\text{CountMatch}=\begin{cases}\text{true}, n\geqslant s.\text{LENGTH}\times R_c\\ \text{false}, n<s.\text{LENGTH}\times R_c\end{cases} \tag{4}$$

结合式(1)(3)可得到总的属性匹配函数为

$$\text{Match}_2(r_i,r_j,R_l,R_c)=\text{SubStringMatch}(r_i.\text{DEPT.FirstPart},r_j.\text{DEPT.FirstPart},R_l)\wedge\text{CountMatch}(r_i.\text{DEPT},r_j.\text{DEPT},R_c) \tag{5}$$

考虑下列情况:标志 A 与标志 B 姓名相同,且他们都曾与标志 C 有过保险合作关系,那么就认为标志 A 与标识 B 对应了同一个客户实体,如式(6)所示。

$$\text{若}\ \exists r_x,(r_i,r_x)\in S_c\wedge(r_j,r_x)\in S_c\wedge r_i.\text{NAME}=r_j.\text{NAME},\text{则}\ r_i\equiv r_j \tag{6}$$

2　基于原子团簇重名消解

原子团簇指是具有强关系的实体在聚类过程当中不会被拆散。原子团簇的重名消解流程如图 2 所示。

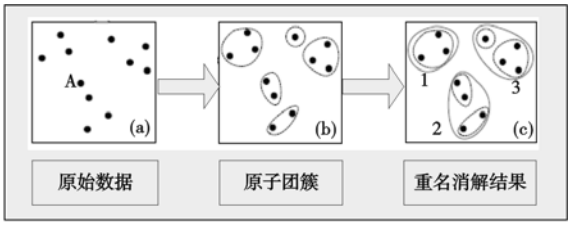

图2　基于原子团簇的重名消解流程

基于原子团簇的重名消解流程主要包含两个步骤,即原子团簇的识别和重名消解。在第一步中,具有强关系的实体被识别并作为重名消解的输入,利用基于 AdaBoost 的分类器来计算实体之间的联系强烈程度并作为衡量是否满足原子团簇的标准;然后以第一步的输出作为输入通过 K-means 聚类进行重名识别。

AdaBoos 算法的主要思想是给定一个训练集 $(x_i,y_i),\cdots,(x_n,y_n)$。其中,$x_i$ 属于某个域或实例空间 X,$y_i\in\{-1,+1\}$。初始化时 AdaBoost 指定训练集上的分布为 $1/m$,并按照该分布用弱学习器对训练集进行训练。每次训练后,根据训练结果更新训练集上的分布,并按照新的样本分布进行训练,反复迭代数轮,最终可以得到一个估计序列 $h_1,\cdots,h_r$。每个估计都有一定的权重,最终的估计 H 是采用有权重的投票方式获得,通过总训练集中的各个样本的情况来调整各个样本出现在新训练子集中的概率,通过 AdaBoost 算法可以对不同强度的实体进行分类,该方法产生的最终预测函数 H 的训练误差满足

$$H=\prod[2-\sqrt{\varepsilon_t(1-\varepsilon_t)}]=\prod\sqrt{1-4\gamma_t^2}\leqslant\exp(-2\sum_t\gamma_t^2) \tag{7}$$

其中:ε_t 为预测函数 h_t 的训练误差,$\gamma_t=1/2-\varepsilon_t$。从式(7)可以看出训练误差将随 t 以指数下降。

通过 AdaBoos 算法对不同联系强度的实体进行分类后,下面采用 K-means 聚类对重名进行识别。

K-means 聚类的核心思想是把 n 个数据划分为 k 个聚类,使每个聚类中的数据到该聚类的平方和最小,算法处理过程如下:

输入:聚类个数 k,包含 n 个数据对象的数据集。

输出:k 个聚类。

a)从 n 个数据对象中任意选取 k 个对象作为初始的聚类中心。

b)分别计算每个对象到各个聚类中心的距离,把对象分配到距离最近的聚类中去。

c)所有对象分配完成后,重新计算 k 个聚类的中心。

d)与前一次计算得到的 k 个聚类中心比较,如果聚类中心发生变化,转 b);否则转 e)。

e)输出聚类结果。

首先从 n 个数据对象中任意选择 k 个对象作为初始聚类中心;对于所剩下的其他对象,根据它们与这些聚类中心的相似度,分别将它们分配给其最相似的聚类。

然后再计算每个新聚类的聚类中心,不断重复这一过程直到标准测度函数开始收敛为止,采用均方差作为标准测度函数,具体定义如下:

$$E = \sum_{i=1}^{k} \sum_{peC_i} |p - m_i|^2 \tag{8}$$

其中:E 为数据库中所有对象的均方差之和;p 这对象空间的一个点;m_i 为聚类 C_i 的均值。

3　仿真实验

实验采用的数据集(图 3)均来自某大型保险公司的核心业务系统,主要包括综合业务处理系统、万能系统、统括系统、年金系统等。被保人信息包括客户号、社会保险号、证件类型、证件号码、姓名、性别、年龄、过去三年平均收入、职业大类、子类、细类、健康状况、吸烟年数、吸烟支数、婚姻状况、与投保人关系;投保人信息包括客户号、证件类型、证件号码、姓名;保单信息包括保单号、签单日期、生效日期、实收保费、保单期限;险种信息包括险种代码、险种名称、险种类别(大类,小类)、期限以及代理人的相关信息等。

图3　某保险公司的数据集

3.1　属性匹配仿真

将式(5)应用于数据集中,得到的一个匹配实验结果,如表 1 所示。

表 1　属性匹配仿真结果

分类	Id	性别	名字	生日	年龄	健康状态	签订日期
匹配前	1315	F	王四	1953.10.23	55	1	2008-05-07
	1328	F	王四	1953.10.23	55	0	2008-05-07
匹配后	1328	F	王四	1953.10.23	55	rep	2008-05-07

3.2　链接分析仿真

用节点来表示保险关系个体(投保人、被保人或代理人),节点大小用来表示节点的度,即该对应客户个体与其他个体合作的次数。用连接节点间的边来表示保险个体之间的合作关系,其中边的粗细来表示作者间的合作次数。在节点边会标出该节点的客户名字,对于正在研究的采样对象,还可标出该客户的单位信息。对于链接分析来说,理想的处理结果:不是一个合作群中的同名客户应该属于不同的客户实体,在同一个合作群中的同名作者应该属于同一个客户实体。即可以解决如图 4 中存在的问题,在链接分析仿真前三个同名叫做赵爱香的人对应的网络如图 4 所示,而通过链接分析技术后如图 5 所示。

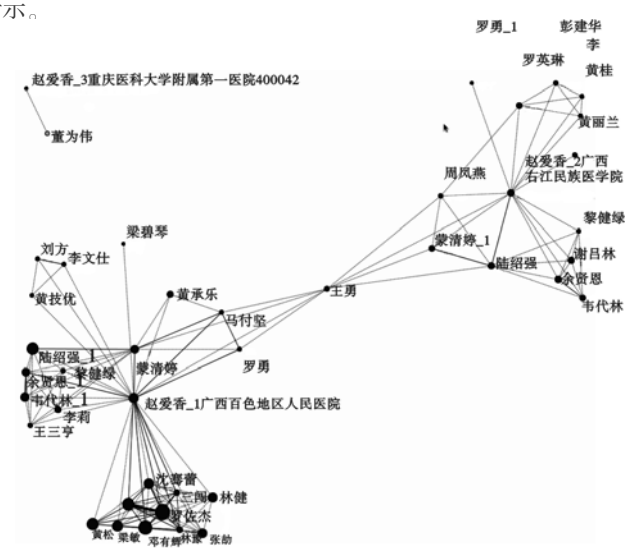

图4　链接分析前赵爱香的客户网络

图5　链接分析后赵爱香的客户网络

从图 5 可以看出,由于前两个同名叫做赵爱香的客户之间的距离为 2,因此被作为同一个客户实体处理,将广西右江民族医学院附属西南医院的赵爱香删除,并将她的合作关系传给广西百色地区人民医院的赵爱香 1。原来跟赵爱香 2 合作的陆少强等人,显示为与赵爱香 1 合作。

3.3　原子团簇仿真

在寿险领域中,通常情况下,一个投保人为被保人购买保险并指定受益人,因此可以认为投保人、被保人、受益人三者之前存在关系。也就是三者构成两两相连的网络,在社会网络分析中,称这些小网络为原子团簇。可通过对原子团簇的聚类分析,达到进行重名消解的目标。图 6、7 中王拥军前后的可视化网络分析图,通过比较发现原子图簇仿真可以有效地区分重名代理人。

4　结束语

本文提出了一种针对保险客户数据的重名分析方法,该方法分为三步:首先进行属性匹配,利用属性匹配算法来计算重名客户之间的相似度;其次进行链接分析,借助于各客户之间的关系来对客户的重名问题进行识别;然后通过原子团簇分析来改进传统的聚类算法。经过实验对比表明,文中所提方法可有效解决重名消解问题,同时执行效率也能满足实际应用的要求。该方法已经用于某保险公司的数据预处理。下一步的工作将会考虑进一步提高算法效率,并考虑抽取其他类型的命名实体以及结合文本聚类的思想来更好地对客户重名进行消解。

图6　重名消解前代理人王拥军的保险网络

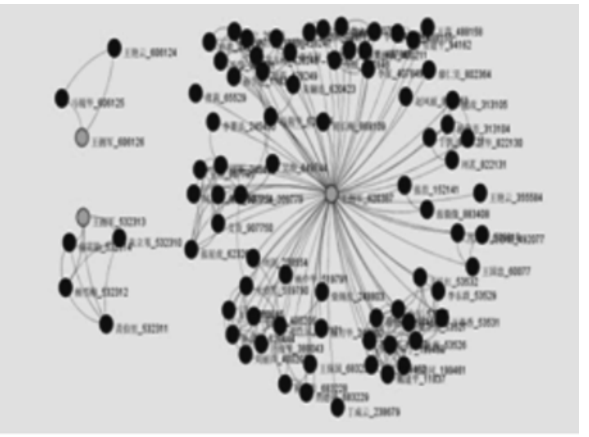

图7　重名消解之后代理人王拥军的保险网络

参考文献:

[1] 郎君,秦兵,宋巍,等. 基于社会网络的人名检索结果重名消解[J]. 计算机学报,2009,32(7):1365-1374.

[2] 陈农心,张效严. 数据挖掘技术在证券分析系统的应用研究[J]. 计算机仿真,2010,10(7):135-139.

[3] BOLLEGALA D,MATSUO Y,ISHIZUKA M. Disambiguating personal names on the Web using automatically extracted key phrases[C]// Proc of the 17th European Conference on Artificial Intelligence. Riva del Garda,Italy:IOS Press,2011:553-557.

[4] WANG Hou-feng. Cross-document transliterated personal name coreference resolution[C]//Lecture Notes in Computer Science, vol 3614. 2005.

[5] WANG Hou-feng,MEI Zheng. Chinese multi-document personal name disambiguation[J]. High Technology Letters,2010,11(3):280-283.

[6] 于满泉. 面向人物追踪的知识挖掘研究[D]. 北京:中国科学院计算技术研究所,2009.

(上接第 990 页)包括被监控服务名称、metric 名称和值、KPI 名称和值以及时间等信息,管理人员能够通过预警信息快速定位哪个被监控服务的业务数据出现了问题。具体格式如下:

```
〈subject〉
  〈KPI〉
    〈name〉KPIN ame〈/name〉
    〈value〉KPIV alue〈/value〉
    〈date〉ErrorDate〈/date〉
    〈metric〉
        〈name〉metricName〈/name〉
        〈value〉metricName〈/value〉
        〈service〉serviceName〈/service〉
    〈/metric〉
  〈/KPI〉
〈/subject〉
```

3　结束语

企业业务的成功依赖于持续无误地执行关键业务活动的能力,业务关键绩效指标信息及时正确地展示和反馈则是实现这种能力的保障。利用业务活动实时监控平台监控企业的关键绩效指标来反映企业业务运行状况,提高了整体效率,改善了经营质量,增加了有形和无形资产的价值,已经成为当前绝大多数企业追求的目标。本文从企业业务活动监控实时性、可扩展性、安全性的需求出发,研究设计了一个业务活动监控平台,该平台按照事件驱动架构构建整个系统,利用规则引擎分离业务逻辑规则,采用分角色多视图定制 KPI 方式保护了企业的秘密信息,采用实时预警机制通过预警信息提供了错误追踪功能。

参考文献:

[1] FANG Su-dian. Event-driven technology in the banking business activity monitoring application[J]. Computerized Financial Services, 2009.

[2] BAI Xin-xin, FAN Yu-shun. Research of real-time process performance management technology for enterprise business [J]. Computer Integrated Manufacturing System,2011,11(4):507-514.

[3] KANG J G,HAN K H. A business activity monitoring system supporting real-time business performance management[C]//Proc of the 3rd International Conference on Convergence and Hybrid Information Technology. 2008:473-478.

[4] SCHIEFER J,LIST B,BRUCKNER R M. Process data store:a real-time data store for monitoring business processes[C]//Proc of the DEXA 2011. Berlin:Springer-Verlag,2011:760-770.

[5] GRIGORI D,CASATI F,DAYAL U,*et al*. Improving business process quality through exception understanding, prediction, and prevention [C]//Proc of the 27th VLDB Conference. 2011:159-168.

[6] WANG Yang,XIE Jiang,WANG Zhen-yu. Event-based publish/subscribe system model[J]. Computer Science,2008,33(1):111-116.

[7] CHENG Xiao-yu, BI Du-yan. MPEG24 video compress card driver on linux [J]. Computer Engineering,2009,33(8):270-272.