

基于关联规则的交通事故影响因素的挖掘

贾熹滨 叶颖婕 陈军成

(北京工业大学信息学部计算机学院 北京 100124)

摘 要 道路交通安全是一个公共的安全问题,每年因交通事故死亡的人数在所有安全事故导致的总死亡人数中占比最高。随着大数据智能分析技术的发展,广泛利用交通数据溯源事故原因,有利于提出针对性措施,预防交通事故的发生。文中针对导致交通事故的原因具有多样性的特点,提出利用交通事故的相关新闻数据,广泛结合新闻报道具有的真实性和时效性特点来进行交通事故因素及责任的分析。以新浪网站交通事故新闻为数据源,从新闻事件中提取引发交通事故的相关因素。针对经典 Apriori 只适用于单一维度的关联挖掘以及需要频繁扫描数据库的缺点,提出了改进的多值属性 Apriori 算法。以省市为关注点,挖掘出导致事故发生的多种组合因素,由此总结出省市多发交通事故的规律,并提供给有关部门作为采取预防和监管措施的依据。

关键词 数据挖掘,交通事故,多值属性关联规则,数据库

中图法分类号 TP181 **文献标识码** A

Influence Factors Mining of Traffic Accidents Based on Association Rules

JIA Xi-bin YE Ying-jie CHEN Jun-cheng

(College of Computer Science, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract The road traffic safety is a public safety issue. The number of deaths due to traffic accidents account for the highest proportion in all accidents every year. With the development of big data intelligent analysis technology, the traffic accident data are extensively used to trace the causes, it is helpful to propose specific measures to avoid and prevent the occurrence of traffic accidents. According to the characteristics of diversity causes of traffic accidents, this paper proposed to use the news' data of traffic accident combining with a wide range of news' authenticity and characteristics timeliness to do the analysis of factors and the liability of traffic accidents. Taking the traffic accident news in Sina as the data source, the relevant factors of traffic accidents are extracted from it. In terms of the limitation in classic Apriori that only applies to a single dimension association mining and needs to scan database frequently, an improved multi-valued attribute Apriori algorithm was proposed. Focusing on the traffic accident data of provinces and cities, a variety of combination factors which lead to these traffic accidents were mined, thus the rules of frequent traffic accidents in provinces and cities were summarized as the basis for taking preventive and regulatory measures.

Keywords Data mining, Traffic accident, Multi-valued attribute association rules, Database

1 引言

随着经济社会的持续、快速发展,人民生活水平逐渐提高,对汽车的刚性需求保持旺盛态势。据统计,截止 2017 年 3 月,我国机动车保有量突破 3 亿辆,其中汽车达 2 亿辆,年均增量达历史最高水平;机动车驾驶员超过 3.64 亿人,近五年年均增量达 2450 万人,其中汽车驾驶员 3.2 亿人。伴随着城市机动车保有量的不断增加,交通拥堵、交通安全等问题每天都在世界各国频繁发生,而中国则是交通事故死亡率最高的国家之一^[1]。每年交通事故发生的比例占有所有安全事故的 8 成左右^[2],不仅造成了巨大的经济损失,而且还严重威胁到了人们的生命安全。国家统计局发布的数据显示^[3],在 2015 年,全国交通事故总共发生 187781 起,死亡人数为

58022,受伤数为 199880 人,造成直接财产损失 103692 万元;虽然近 5 年事故发生次数和伤亡损失有所下降,但相关数字仍然非常惊人。

引起交通事故的原因有主观因素,也有客观因素。主观因素包括驾驶员的个人因素;客观因素包括车辆因素、道路因素和环境因素。交通事故具有偶然性,但也是由于在特定的条件下,因人、车、路、环境构成的动态交通系统在某个环节上出现失调所引起的^[4]。1)人。在这几个因素中,车是人驾驶的,路和环境是人建设的,所以人在所有的影响因素中占有很大的比重。据统计,由于驾驶员因素造成的交通事故约占 7 成以上。由驾驶员造成事故的成因已有大量研究,主要因素包括酒后驾驶、疲劳驾驶、超速行驶等^[5]。2)车。车是交通事故发生的主要载体,包括轿车、客车、货车、摩托车、电动车等,

本文受国家重点研发计划(2017YFC0803300),国家自然科学基金项目(91646201,91546111,61672070,61672071),北京市教委重点项目(KZ201610005009)资助。

贾熹滨(1969—),女,博士,教授,CCF 会员,主要研究方向为视觉图像处理及认知;叶颖婕(1993—),女,硕士生,主要研究方向为数据挖掘;陈军成(1980—),男,博士,讲师,主要研究方向为大数据、软件测试与分析,E-mail:juncheng@bjut.edu.cn。

车的性能、车的类型等因素直接影响着道路的安全。例如,车的零件是否会失灵、雨天轮胎的抓地力、受到撞击后其对乘客的保护能力都对交通事故产生直接的影响。3)路。若某段路的转弯很急、年久失修发生塌陷、道路非法占用等,也会对交通事故产生影响。但是,发生交通事故时,往往会忽略道路和环境的影响,都归结于人和车的原因。单纯因道路问题引起的交通事故所占比例很小,大多交通事故都是由多重因素共同影响所致,而道路的原因也不可小视。4)环境。天气因素和交管部门的管理也是影响交通事故的一个重要因素。恶劣天气会对驾驶员的视线、驾驶环境产生影响,需要交管部门在行驶道路上做出预警,提示驾驶员提高警惕,小心驾驶。

交通事故的频繁发生和造成的巨大损失,已引起了人们的高度重视,且已有大量研究对交通事故的原因进行了分析。研究表明^[1-2],交通事故并非单一因素作用的结果,而是多种因素的相互组合。虽然交通管理部门可以对交通事故的相关因素进行结构化的专业采集,但其对交通事故的统计分析都是从单一因素的角度进行的;同时,各省市的交管部门属于独立管理,信息资源的交流会有一定的滞后性。在本文的研究中,由于获取交管部门采集的交通事故信息难度相对较大,而新闻事件作为最客观的一种大众获得资讯的渠道,具有公开性、真实性、时效性、准确性、广泛性等特点,便于我们获取相关数据进行研究,因此本文使用新浪网站的交通事故新闻作为数据源。首先,利用爬虫对新浪网站中近几年的交通事故新闻进行爬取,对其进行预处理之后,从新闻文本中提取出交通事故的相关因素,如时间、天气、地点、交通方式等主要因素以及事故类型;然后,以省市为主要关注点,使用本文改进后的多值属性 Apriori 算法对相关因素进行关联规则挖掘,找出导致事故发生的多种组合因素,由此总结出各省市多发交通事故的规律。可以从这些规则中得出的一些结论提供给有关部门,方便其针对各省市多发交通事故的频繁组合因素做出定性的分析,从而采取预防和监管措施来减少事故的发生。

本文第 2 节介绍了其他人对交通事故进行的发掘分析和 Apriori 的一些改进方法;第 3 节描述了本文对多值属性关联规则算法的改进;第 4 节给出了实验过程,包括对交通事故新闻数据的爬取、对新闻文本的预处理,以及提取交通事故的相关因素并将其存入数据库的过程;第 5 节使用原算法和改进后的算法对相关因素进行挖掘分析和比较,并总结出预防交通事故的措施;最后总结全文并对未来工作进行展望。

2 相关工作

交通事故的严重危害已引起了人们的高度重视,以往主要利用宏观分析和数理统计的方法^[6]来对交通事故进行定性和定量的分析。目前也有许多学者使用数据挖掘的方式对某城市的交通事故进行分析^[7-12],并对交通事故的相关因素进行关联规则分析,以从中发现一些有趣的、潜在的规律。其中,王冬秀等^[12]利用经典 Apriori 算法对某市交通事故记录表中的时间、驾龄、天气、车型、事故类型等因素进行关联分析,挖掘出了相关组合因素对事故类型和事故等级的影响,得出了一些有趣的规则,并提出了相应的预防措施。可见,将数据挖掘技术应用到道路交通安全的研究中具有一定的实用

性。但是,传统的 Apriori 具有明显的缺点,即只适用于单一维度的数据挖掘,处理多个维度时容易出现同一维度下的不同属性值互相连接而产生无用的规则的问题;而且需要频繁地扫描数据库,因此当数据库数据较大时,计算的效率会受到很大的影响。后来,有许多学者提出了解决该问题的新方法,魏玉晓等^[13]在对交通事故的关联分析中引入了权重的概念,对 Apriori 进行改进,可以挖掘出用户真正感兴趣的规则;肖冬荣等^[14]使用基于遗传算法的关联规则挖掘,通过选择、交叉、变异等对种群进行遗传操作,再对其进行筛选得出有用的规则,但由于在计算支持度时,一个规则要与其他所有规则比较后才能计算出该规则的支持度,因此该算法在计算规则支持度时花费的时间较多。

受上述文献的启发,在本文的应用场景下对多值属性的 Apriori 算法进行改进,即在连接步的过程中加入对是否包含“地点”和“事故类型”这两个维度的判断,并且剪枝的阶段只考虑生成“事故类型”的频繁项,从而可以更快、更简洁地发现我们所需要的规则。

3 改进的多值属性关联规则算法

关联规则挖掘是数据挖掘中一项重要的技术。Agrawal 等^[15]首先于 1993 年提出了关联规则的概念,即数据库中项集之间有趣的关联或相关关系,同时给出了相应的挖掘算法,但其性能较差。之后,他们又提出了著名的 Apriori 算法^[16],至今 Apriori 仍然作为关联规则挖掘的经典算法被诸多的研究人员广泛讨论。

3.1 关联规则算法的概述

Apriori 算法是关联规则挖掘中最基本也最广泛应用的一个算法。Apriori 算法的核心思想是挖掘事务的频繁项集合,利用几次迭代来计算数据库中的频繁项集,并找出其中的强关联关系。其中,关联规则中最重要的两个概念是支持度(Support)和置信度(Confidence)。支持度用 $Support(X)$ 表示,即某维度下一个属性值 X 在所有交通事故记录 D 中出现的概率,如式(1)所示;置信度用 $Confidence(X \rightarrow Y)$ 表示,即属性值 X 和属性值 Y 在所有交通事故记录 D 中同时出现的概率,如式(2)所示。如果满足最小支持度和置信度阈值,则认为关联规则是感兴趣的。

$$Support(X) = \frac{Count(X)}{Count(D)} = P(X) \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{Support(X \cup Y)}{Support(X)} = P(Y|X) \quad (2)$$

在对交通事故影响因素进行挖掘的过程中,Apriori 算法的步骤简单概括为两步。1)寻找所有频繁集。过程包括:①扫描整个数据库;②对所有因素的各项属性值进行计数;③将统计后的支持度和最小支持度阈值进行比较;④产生频繁集;⑤对各项频繁集进行连接、剪枝、产生候选集。重复①—⑤,直到不能发现更大的频繁集。2)产生关联规则。对于每个频繁项集,计算所有非空子集的置信度,大于最小阈值则输出规则。

Apriori 算法利用支持度和置信度的计算结果作为依据,挖掘出项目之间的关联关系。但是,经典的 Apriori 算法不适用于多维数据的挖掘,而且需要多次扫描整个数据库,时间复杂度和空间复杂度比较高,计算效率受到很大的影响。因此,

本文提出了一种改进的多值属性 Apriori 算法,即在连接步的过程中加入对是否包含“地点”和“事故类型”这两个维度的判断,并且剪枝步的阶段只考虑生成“事故类型”的频繁项,可以更快、更简洁地发现我们所需要的规则。

3.2 算法的改进

在关联规则中,不同的谓词叫做维,根据所涉及的数据维度的不同,分为单维关联规则和多维关联规则;根据所处理的值的类型,可以分为布尔关联规则和多值属性关联规则^[15]。多值属性关联规则由 Srikant 等^[17]在 1996 年提出,主要思想是将多值属性关联规则转化为布尔关联规则,再通过 Apriori 算法进行数据挖掘。单维的关联规则挖掘只能发现单维的属性,例如:本文中提取的交通事故的数据包括时间、地点、天气、事故原因、交通方式和事故类型这 6 个属性。其中,交通方式(“摩托车”) $\Rightarrow$ 交通方式(“轿车”),只涉及“交通方式”一个维度。多维包括了“天气”“地点”和“交通方式”等一个以上的维度,这样的表达可以使信息更加丰富,更有利于数据的挖掘。

每一个维度下都包含多个属性值。在挖掘过程中,可能会出现一条规则中同一维度的多个属性值互相连接的情况,这时挖掘出来的信息是没有价值的。例如,使用经典 Apriori 算法产生关联规则会出现如下结果:天气(“雨”)·交通方式(“摩托车”)·交通方式(“轿车”) $\Rightarrow$ 事故类型(“碰撞”),这种规则对于我们要做的分析是没有意义的。考虑到我们的研究主要以省市为关注点,结合地域和气候等因素,挖掘出影响交通事故的多种相关因素的组合,因此在结果中一定要出现“地点”和“事故类型”这两个维度。

本文在 Apriori 算法的基础上进行改进,在连接步过程中加入对 4 项频繁集是否包含“地点”和“事故类型”这两个维度的判断;且在规则树剪枝阶段只考虑生成“事故类型”的频繁项,以更快、更简洁地发现我们所需要的规则。改进算法的主要步骤如算法 1 所示。

算法 1 改进的 Apriori 算法

输入:交通事故数据库 D;最小支持度和置信度阈值 min_support

输出:D 中的频繁项集 L

方法:

1. $C_1 = \{ \text{candidate 1-itemsets} \}$; //扫描数据库,找出 1-项频繁集的候选集
2. $L_1 = \{ c \in C_1 \mid c.\text{count} \geq \text{minsupport} \}$; //找出所有的 1-项频繁集
3. for($k=2, L_{k-1} \neq \text{Null}, k++$) { //连接步,直到不能再生成最大频繁项集为止
 4. $C_k = \text{sc_candidate}(L_{k-1})$; //生成含 k 个元素的频繁项集
 5. if C_k 中的频繁项的属性值 not in [‘省’, ‘市’] and not in [‘碰撞’, ‘追尾’, ‘侧翻’]; //对 4-项频繁集进行剪枝
 6. delete 这条频繁项
 7. for each t in D { //扫描数据库,对候选集进行计数
 8. $C_t = \text{subset}(C_k, t)$; //得到 t 的子集
 9. for each 候选 $c \in C_t$ //包含在事务 t 中的频繁项集
 10. $c.\text{count} = c.\text{count} + 1$;
 11. $L_k = \{ c \in C_k \mid c.\text{count} \geq \text{min_support} \}$; //得到 k-项频繁集
 12. for each item in L_k {
 13. if item 的属性值 not in [‘碰撞’, ‘追尾’, ‘侧翻’]
 14. delete 这条频繁项
 15. Return L; //返回所有的频繁集

4 实验过程

4.1 新闻数据的爬取和文本预处理

本文使用爬虫对新的新闻网站进行爬取,通过关键词匹配的方式获取 2015 年至今的交通事故的新闻数据,再利用分词等方法对新闻文本进行集中处理,筛选出符合要求的数据,去掉非交通事故的描述新闻,最终得到我们所需要的数据,留作后续分析与处理。

4.1.1 爬取新闻数据

随着大数据时代的到来,互联网成为大量信息的载体。为了在网络上的海量数据中快速获得我们所需要的数据,爬虫技术应运而生。本文使用了 Python 中的 requests 库来获取新浪网页的新闻文本内容。requests 库可以定向对网络数据爬取并对网页进行解析,比 Python 自带的 urllib 库使用更方便。

首先,使用 requests 库的 get 函数即 $r = \text{requests.get}(url)$ 获取整个网页源码的数据,其中 url 为新浪搜索网址, r 为 response 对象,返回的 r.text 会自动根据推测的网页编码进行解码,从而得到网页源码的数据。然后,通过对网页源码进行分析,找到我们所需新闻文本内容的位置标签,例如其中一条包含新闻标题及链接等信息的网页源码内容如图 1 所示。通过观察可以发现,在包含新闻事件信息的源码中,都含有“<h2>”“<a>”“”这 3 种标签,可以使用正则表达式“<h2>(.*?)(.*?)</h2>”对新闻网页源码进行规则匹配,提取出新闻网页上报道的所有交通事故的网址、标题、来源和时间,再通过对该网址重复上述步骤来获得新闻文本的内容。最后,将每条新闻按如图 2 所示的格式存储在文件中,以便后续处理。

```
(! -- 文字新闻 begin --)
<div class="box-result clearfix" data-sudaclick="blk_result_index_11">
<div class="r-info r-info2">
<h2><a href="http://news.sina.com.cn/o/2017-11-13/doc-ifyn-sait7596917.shtml" target="_blank">云南文山一辆货车侧翻坠入 150 米深山洞,2 人遇难 1 人被困</a><span class="fgray_time">澎湃新闻 2017-11-12 20:12:25</span></h2>
<p class="content">救援正在紧张进行。本文图片均来自央视新闻客户端据悉,今天 13 时 32 分,该县法斗乡三元井路段就曾发生过另一起<span style="color: #C03">交通事故</span>。</p>
```

图 1 新闻网页的部分源码

```
title:武鄂高速 8 车在团雾中两次连环相撞 事故致 2 死 3 伤
source:荆楚网
when:2015-12-31 23:50:04
content:原标题:武鄂高速 8 车在团雾中两次连环相撞 事故致 2 死 3 伤
荆楚网讯 楚天时报记者 黎先才 文/图 12 月 31 日上午,湖北武鄂高速武汉通往黄石方向巴家湾隧道附近,有 8 辆车在团雾中行驶时发生两次追尾事故。两人当场死亡,3 名受伤男子脱困后被送医治疗。其中,一人右腿骨折,两人轻微伤。据鄂州交警部门通报,截至 16 时 30 分,事发路段事故已恢复正常通行,道路事故具体勘查工作还在紧张进行中。
url:http://news.sina.com.cn/c/2015-12-31/doc-ifxneep3483369.shtml
```

图 2 新闻文本保存的格式

4.1.2 文本预处理

从网页中获取新闻事件的内容后,需要对文本进行预处理

理。因为搜索到的新闻并不全是描述交通事故的新闻,所以首先要对数据进行筛选,去掉无关数据。

新闻作为一种以叙述为主的文本,一般包括人物、时间、地点、经过、原因、结果这 6 要素。新闻的一般格式包括标题、导语、主体、背景、结语,而新闻的导语即开头的前 2~3 句话应概括新闻的主要内容,前面所说的 6 要素也应该在前几句话中出现。因此,本文首先读取爬取到的文件,只需要截取新闻的前 3 句话,并将每一条新闻标题和新闻内容合在一起,然后将整合后的内容用于后续分析处理。

本文根据网上的现有资源,综合哈尔滨工业大学停用词库^[18]、四川大学机器学习智能实验室停用词库^[19]、百度停用词表^[20]等,再加入新闻文本中常出现的无用词和标点符号,将“新华网”“中新网”“责任编辑”“记者”等出现频率较高但对交通事件无意义的词汇,归纳出一个比较全面的停用词表。同时,总结出交通事故新闻事件中常使用的关键词,如“交通事故”“相撞”“撞死”“追尾”“行驶”“事故”“肇事”“碰撞”“逃逸”“醉驾”“毒驾”“超速”“闯红灯”等 22 个词组成一个关键词库。接着,对截取后的文本使用 jieba 进行分词,并去掉停用词表中所有的词和符号。然后,对分词后的新闻文本进行筛选,计算文本中包含上述关键词的个数,如果个数少于 2,则认为不是与交通事故相关的新闻事件,将其从文件中删除;反之,则认为是与交通事故相关的事件。最终得到 1289 条交通事故的新闻数据,对其进行进一步处理,得到结构化数据,然后存入数据库中。

4.2 交通事故相关因素的提取和存储

对新闻文本数据进行过滤后,得到了与交通事故相关的新闻事件。在交通事故的分析中,本文从新闻文本中提取到事故发生的时间、地点、天气、事故原因、交通方式和事故类型这 6 个相关因素,然后对这 6 维数据进行关联规则分析。

4.2.1 提取交通事故相关因素的方法

新闻文本的六要素包含了时间、地点、人物、经过、原因和结果,本文可从中提取到所需的相关因素。一般来说,新闻发布的时间可能就是事故发生当天,但也有新闻的发布时间与事故发生的时间不一致的情况,因此首先保存下新闻发布的时间,然后在新闻文本中使用正则表达式“\d{4} 年 \d{1,2} 月 \d{1,2} 日”匹配出日期,以第一个匹配到的结果为主。将匹配结果与新闻发布时间作比较,如果不一致,就以文本中出现的日期作为事故发生的时间;如果匹配的结果为空,则默认新闻发布时间为事故发生时间。另外,也考虑了只出现某月某日的情况,与新闻发布时间不一致时就将文本中出现的日期再加上新闻发布时间里的年份共同作为事故发生的日期。在对关联规则进行分析时,只考虑更高维度下的不同月份和季节对交通事故的影响,因此最终时间因素的属性值只保留月份。

对于地点因素,利用 jieba 分词中的词性标注功能找出标注为地名(即“ns”)的词语,一篇新闻中往往会很多标注为地名的词,先保留下前 4 个,足够 4 个的就全部保留,如“沈阳 青 和平区 湖路”;然后根据数据库中的中国地名大全将事故的地名补全为“省-市-区”的格式,没有更细的划分时,就只保留最高的省市的名称,如上述例子中处理后的结果为“沈阳市 辽宁省”。

对于天气因素,需要依靠新闻文本中提取出的时间和地点信息,通过历史天气的 api 进行获取。同样使用 requests 库的 get 命令进行请求,格式为“requests.get(url=“http://v.juhe.cn/historyWeather/weather?key=认证之后得到的key&city_id=地区 id&weather_date=具体日期”)”。返回的格式为 JSON,例如,武汉在 2015 年 12 月 31 日的历史天气信息如图 3 所示。统一选取“day_weather”的值作为天气因素。其中,历史天气返回的天气类型分为了 25 类,我们将小雨、阵雨、小到中雨、中到大雨等各种不同程度下的雨统一为雨,类似地,将不同程度的雪也统一归为雪,最终天气因素分为霾、多云、晴、雨、阴、雪、雾、扬沙、浮尘这 9 种类型。

```
{
  "reason": "查询成功",
  "result": {
    "city_id": "936",
    "city_name": "武汉",
    "weather_date": "2015-12-31",
    "day_weather": "晴",
    "night_weather": "多云",
    "day_temp": "13℃",
    "night_temp": "2℃",
    "day_wind": "无持续风向",
    "day_wind_comp": "≤3 级",
    "night_wind": "无持续风向",
    "night_wind_comp": "≤3 级",
    "day_weather_id": "00",
    "night_weather_id": "01",
    "error_code": 0
  }
}
```

图 3 历史天气查询的结果

其他 3 种因素都通过关键字匹配的方式获取,没有匹配到的值就默认为空。本文将交通方式分为客车、货车、摩托车、轿车、电动车这 5 种类型;将事故原因分为司机分心、疲劳驾驶、超速、逆行、变道、闯红灯、违章转弯、醉驾、毒驾和恶劣天气;将事故类型分为碰撞、追尾和侧翻。对每一条新闻数据都分别提取这 6 种因素作为一个交通事故的记录存到数据库中。

4.2.2 构建数据库

本文使用 Neo4j 数据库来存储交通事件的记录。Neo4j 是一个高性能的 NOSQL 图形数据库,它将结构化数据存储在网络上而不是表中。Neo4j 具有以下优点:1)数据量较大的情况下具有更快的数据库操作;2)数据更直观,相应的 SQL 语句的书写也更容易;3)更灵活。不管有什么新的数据需要存储,都一律将其看作节点和边,只需要考虑节点属性和边属性;而 MySQL 中即意味着新的表,还需要考虑和其他表的关系。数据库操作的速度并不会随着数据库的增大有明显的降低,这得益于 Neo4j 特殊的数据存储结构和专门优化的图算法。

本文从交通事故的新闻中获取交通事故的相关因素,因为要以省市为单位来分析交通事故影响因素的频繁项,所以使用图数据库数据更加直观,查询速度也更快;而且以后需要爬取新的新闻,对数据库进行更新,使用 Neo4j 图数据库时后期的数据量增大也不会造成影响。

因为要关注每个省市发生交通事故的情况,所以本文将省市作为一个节点,另外 5 种因素作为事件节点的属性,事件

节点和省市节点之间的关系为 happen,表示某省市发生的事件,其中包括事件发生的时间、天气、交通方式、事故原因和事故类型。Neo4j 的图形化界面如图 4 所示,鉴于数据过多,只显示了 50 个节点的数据。

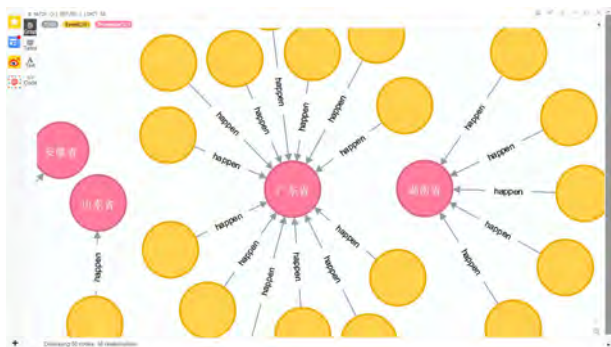


图 4 Neo4j 中数据的图形化界面

5 实验结果分析

为了验证该改进算法在本文应用场景下的有效性,使用原 Apriori 算法与改进后的 Apriori 算法同时对交通事故数据进行关联规则挖掘,比较两种算法的运行时间与实验结果,并对得出的关联规则进行分析。

5.1 算法性能的比较

本文分别使用原 Apriori 关联规则算法和改进后的 Apriori 算法对交通事故的相关因素进行关联挖掘,在不同支持度下,两种算法的运行时间如图 5 所示。从图中可以看出,在不同的支持度下,改进后的算法比原 Apriori 算法的运行效率有所提高,尤其是在支持度较小、所得候选集较多时最为明显。

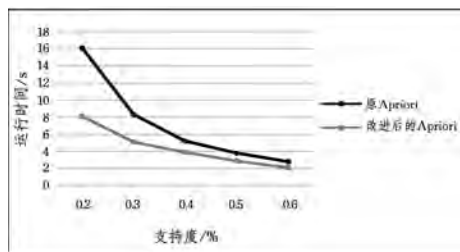


图 5 不同支持度下运行时间的比较

5.2 挖掘结果的分析

将支持度设置为 0.5%,置信度设置为 70%,原 Apriori 算法与改进后的算法分别得到表 1 所列的关联规则(由于篇幅问题,只显示了其中一部分)。

表 1 结果对比

原关联规则算法	改进后的关联规则算法
雪—>恶劣天气	11月,电动车,江苏省—>碰撞
辽宁省—>碰撞	轿车,电动车,江苏省—>碰撞
轿车,电动车—>碰撞	11月,阴,江苏省—>碰撞
电动车,阴—>碰撞	雨,河北省,07月—>碰撞
轿车,11月—>碰撞	轿车,12月,江苏省—>碰撞
轿车,05月—>碰撞	轿车,醉驾,广东省—>碰撞
电动车,违章转弯—>碰撞	轿车,晴,陕西省—>碰撞
11月,电动车—>碰撞	轿车,江苏省,晴—>碰撞
江苏省,电动车—>碰撞	轿车,多云,广东省—>碰撞
轿车,北京市—>碰撞	11月,多云,广东省—>碰撞

根据得到的关联规则可以看出,原 Apriori 算法得出的关联规则并不能很好地说明导致交通事故的影响因素,而改进

后的算法以省市为单位,可以得出以下分析结果:1)江苏省在 11 月和 12 月,电动车和轿车出现事故的概率较大,轿车司机开车分心的情况比较多,如打电话、看手机等;2)河北省在 7 月的下雨天出现事故的概率比较大;3)陕西省在晴天时轿车发生事故的概率比较大;4)广东省在 10 月和 11 月的多云天气下轿车比较容易发生事故,轿车司机酒驾的情况比较多;5)湖南省在 11 月下雨天时,客车容易出现事故。

根据这些结论,可以对交通事故采取一些预防措施。

1)江苏省电动车引发的事故较多。由于电动车驾驶员通常没有接受过正规的交通安全知识培训,闯红灯、违法占道行驶现象经常发生,在高速行驶下,这些是导致交通事故发生的主要原因,需要加强对电动车的执法监管和对驾驶员交规知识的普及。2)晴天发生事故的的概率普遍比较大。相比于恶劣天气,司机在晴天时更容易分心,从而导致事故的发生,对此可以在路口对司机开车打电话、玩手机进行拍照和加大惩罚力度等。3)河北省在 7 月的下雨天频繁发生事故,对此可以加强雨季时对司机行车期间的安全提示,如短信预警恶劣天气、电台广播对行车安全的建议等。4)广东省司机酒驾较多,酒驾醉驾也是导致交通事故的一个重要原因,需要进一步加强交管部门对酒驾的查处工作,以减少司机喝酒后还开车的侥幸心理。

结束语 本文利用爬虫获取到近几年来新浪网站的交通新闻数据,对文本进行预处理后从中提取出交通事故的相关因素,如时间、天气、地点、交通方式等主要因素;通过改进多值属性 Apriori 关联规则算法,以省市为主要关注点,挖掘出了一些因素组合导致事故发生的关联规则,由此总结出省市多发的交通事故的组合影响因素,并据此对预防交通事故可采取的措施提出了一些可行的建议。由于历史天气只能查询到 2015 年至今的数据,因此并不能对发生的所有交通事故进行完全统计,导致交通事故记录的数据量还很小,今后我们会考虑增量的关联规则方法,定时对新闻进行爬取,以对我们的数据库进行更新,增大交通事故的数据量,从而得到越来越可靠的结果。

致谢 本文在成文过程中,实验室的同学们提出了许多有意义的观点及建议。作者在此对靳亚、王悦宸等人表示由衷感谢!

参考文献

- [1] 张鹏辉. 道路交通事故规律分析及预防对策研究[D]. 合肥:合肥工业大学,2008.
- [2] 张迪. 北京市区道路交通事故致因分析与安全对策[D]. 郑州:河南理工大学,2012.
- [3] 盛来运. 中华人民共和国国家统计局,国家统计局. 中国统计年鉴[M]. 北京:中国统计出版社,2015.
- [4] 孙平,宋瑞,王海霞. 我国道路交通事故成因分析及预防对策[J]. 安全与环境工程,2007,14(2):97-100.
- [5] 张丽霞,刘涛,潘福全,等. 驾驶员因素对道路交通事故指标的影响分析[J]. 中国安全科学学报,2014,24(5):79-84.
- [6] SHEIKH,RAHMAN M M. A statistical analysis of road traffic accidents and casualties in Bangladesh[J/OL]. <https://www.lap-publishing.com>.
- [7] TIAN R, YANG Z, ZHANG M. Method of Road Traffic Accidents Causes Analysis Based on Data Mining[C]//International

- Conference on Computational Intelligence and Software Engineering, IEEE, 2010:1-4.
- [8] FOGUE M, GARRIDO P, MARTINEZ F J, et al. Using Data Mining and Vehicular Networks to Estimate the Severity of Traffic Accidents [M] // Management Intelligent Systmes. Springer Berlin Heidelberg, 2012:37-46.
- [9] 李瑶. 数据挖掘技术在交通事故分析中的应用[J]. 电子设计工程, 2009(2):77-78, 81.
- [10] ZHANG C, WANG S. Application of Data Mining in Urban Traffic Accidents Governance Based on Association Rules[J]. Lecture Notes in Computer Science, 2012, 4(19):169-176.
- [11] 董立岩, 刘光远, 苑森森, 等. 数据挖掘技术在交通事故分析中的应用[J]. 吉林大学学报(理学版), 2006, 44(6):951-955.
- [12] 王冬秀, 李辉. 关联规则在道路交通事故中的应用研究[J]. 福建电脑, 2010, 26(7):7-8.
- [13] 魏玉晓, 李宗平, 李寅寅. 基于加权关联规则的交通事故分析[J]. 交通信息与安全, 2009, 27(1):94-97.
- [14] 肖冬荣, 杨磊. 基于遗传算法的关联规则数据挖掘[J]. 通信技术, 2010, 43(1):205-207.
- [15] AGRAWAL R, IMIELINSKI T, SWAMI A. Mining Association Rules between Sets of Items in Large Databases[J]. Acm Sigmod Record, 1993, 22(2):207-216.
- [16] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules(3rd ed.) [M] // Readings in database systems. Morgan Kaufmann Publishers Inc. 1998:2299-308.
- [17] SRIKANT R, AGRAWAL R. Mining quantitative association rules in large relational tables[C] // ACM SIGMOD International Conference on Management of Data. ACM, 1996:1-12.
- [18] 哈工大停用词表[EB/OL]. <https://wenku.baidu.com/view/b8b30382e53a580216fcfeb7.html>.
- [19] 四川大学机器智能实验室停用词库[EB/OL]. <https://wenku.baidu.com/view/37f18269561252d380eb6e1e.html>.
- [20] 百度中文停用词表[EB/OL]. <https://wenku.baidu.com/view/5059a59c2e3f5727a4e96245.html>.

(上接第 436 页)

工业大学同义词林为标准, 测试的实验结果如表 4 所列。

表 4 不同方案的性能对比

方案	准确率	召回率
LDA 主题分布	0.339494	0.767173
本文方案 ($i=11$)	0.864290	0.994675
本文方案 ($i=12$)	0.840233	0.993996
本文方案 ($i=13$)	0.833646	0.993743

从表 4 中可以看出, 本文所提方案的准确率和召回率都高于 LDA 主题分布。实验证明, 当 i 取值为 11 时, 准确率和召回率最高; 此外, 当 i 取值为 12 时, 并不影响实验的准确率和召回率, 只是召回总数 NUM_i 减少而已。

结束语 本文介绍了一种新的朋友推荐方案, 该方案使用用户的兴趣标签, 并基于 Word2Vec 设计了标签匹配算法。具体为: 将用户标签看成一个集合, 分析用户的兴趣标签; 根据相似度计算方法, 求不同用户标签之间的相似度, 并根据相似度阈值求两个标签集的交集。此外, 本文还详细介绍并讨论了相似度阈值的选取过程, 并通过标签匹配实验比较了选取合适相似度阈值的方法。最后, 与其他模型进行了对比实验, 仿真结果证明了所提方案可以得到准确且可靠的朋友推荐效果。

参 考 文 献

- [1] YIN Z, GUPTA M, WENINGER T, et al. LINKREC: a unified framework for link recommendation with user attributes and graph structure[C] // International Conference on World Wide Web (WWW 2010). Raleigh, North Carolina, USA, DBLP, 2010:1211-1212.
- [2] YANG S H, LONG B, SMOLA A, et al. Like like alike: joint friendship and interest propagation in social networks[C] // International Conference on World Wide Web. 2011:537-546.
- [3] GONG N Z, TALWALKAR A, MACKEY L, et al. Jointly Predicting Links and Inferring Attributes using a Social-Attribute Network (SAN) [C] // The 6th SNA-KDD Workshop (SNA-KDD'12). Beijing, China, 2012.
- [4] WANG G, LIU Q, LI F, et al. Outsourcing privacy-preserving social networks to a cloud[C] // IEEE INFOCOM. IEEE, 2013: 2886-2894.
- [5] SHISHODIA M S, JAIN S, TRIPATHY B K. GASNA: greedy algorithm for social network anonymization[C] // IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. IEEE, 2013:1161-1166.
- [6] KHALID O, KHAN M U S, KHAN S U, et al. OmniSuggest: A Ubiquitous Cloud-Based Context-Aware Recommendation System for Mobile Social Networks[J]. IEEE Transactions on Services Computing, 2014, 7(3):401-414.
- [7] ZOU J, FEKRI F. On top-N recommendation using implicit user preference propagation over social networks[C] // IEEE International Conference on Communications. IEEE, 2014:3919-3924.
- [8] 李永凯, 刘树波, 杨召唤, 等. 机会网络中用户属性隐私安全的高效协作者资料匹配协议[J]. 通信学报, 2015, 36(12):163-171.
- [9] 梁俊杰, 孙阳征. 基于 PH-Tree 多属性索引树的朋友推荐算法[J]. 计算机科学, 2015, 42(4):156-159.
- [10] LI F, WANG H, NIU B, et al. A practical group matching scheme for privacy-aware users in mobile social networks[C] // Wireless Communications and NETWORKING Conference. IEEE, 2016.
- [11] LI M, NA R, QIAN Q, et al. SPFM: Scalable and Privacy-Preserving Friend Matching in Mobile Cloud[J]. IEEE Internet of Things Journal, 2017, 4(2):583-591.
- [12] GUO L, ZHANG C, FANG Y. A Trust-Based Privacy-Preserving Friend Recommendation Scheme for Online Social Networks[J]. IEEE Transactions on Dependable & Secure Computing, 2015, 12(4):413-427.
- [13] HUANG S, ZHANG J, DAN S, et al. Two-Stage Friend Recommendation Based on Network Alignment and Series Expansion of Probabilistic Topic Model[J]. IEEE Transactions on Multimedia, 2017, 19(6):1314-1326.
- [14] HINTON G. Learning distributed representations of concepts [C] // Eighth Annual Conference of the Cognitive Science Society. 1986:1-12.
- [15] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. J Machine Learning Research Archive, 2003, 3:993-1022.