基于数据挖掘的个人征信系统异常查询实时监测模型及其应用

姚前，谢华美，景志刚，胡青青，司恩哲
中国人民银行征信中心，北京 100031

摘要
选择个人征信系统最新36个月9亿条查询记录，根据用户查询行为的不同波动特征进行了模型细分，探讨了4种异常查询实时监测模型。结果表明，基于数据挖掘的个人征信系统异常查询实时监测模型应用于个人查询量预测是可行的，且效果良好。该模型的成功上线和不断修正，将对个人征信系统的违规查询行为产生威慑作用，倒逼查询机构加强内部管理，合法使用信用信息，以保障信息主体的权益，促进征信市场健康发展。

关键词
数据挖掘；个人征信系统；异常查询；违规查询；实时监测

中图分类号：TP399　　文献标识码：A　　doi: 10.11959/j.issn.2096-0271.2016044

Real-time data-mining-based anomaly inquiry monitoring model of personal credit reference system and it's application

YAO Qian, XIE Huamei, JING Zhigang, HU Qingqing, SI Enzhe
Credit Reference Center, the People's Bank of China, Beijing 100031, China

Abstract
The data selected contained 900 million query records in the latest 36 months from the personal credit reference system database. The model was subdivided according to different volatility characteristics of each user's query behavior, and four types of real-time anomaly inquiry monitoring models were discussed and developed. Results indicate that the anomaly inquiry monitoring model is feasible to apply on predicting anomaly query behaviors and showed positive effects. The successful application and constant perfection of the model would definitely exert deterrent effect on illegal query behaviors, force commercial banks to strengthen internal management, protect individual's private information and right, and promote the healthy development of the credit reference market.

Key words
data mining, personal credit reference system, anomaly inquiry, illegal inquiry, real-time monitoring

1 引言

全国集中统一的个人征信系统共收集8.8亿个自然人的信用信息，基本覆盖全国每一个有信用活动的信息主体，其中，个人贷款和信用卡账户信息21.5亿笔，开通查询用户15.9万个，对外提供29.2亿份个人信用报告。作为金融系统重要基础设施的个人征信系统，在提高商业银行风险管理水平、提高审贷效率、拒绝高风险客户、清收不良贷款等方面发挥重要作用。2013年3月15日《征信业管理条例》出台并正式实施，是我国征信业发展史上的一个里程碑，不仅严格规范个人征信业务规则，还要求切实保护个人信用信息。该条例要求信息主体以外的单位或者个人向征信机构查询个人信用报告时，应当取得信息主体本人的书面同意并约定用途。信息使用者应当按照与信息主体约定的用途使用个人信息，不得用作约定以外的用途，不得未经信息主体同意向第三方提供。但违规查询个人信用报告的情况时有发生，为了更好地保护信息主体的权益，维护个人征信系统的客观、公正和权威，急需通过数据挖掘技术，分析查询行为，建立异常查询实时监测模型。

为了能更准确地定位异常查询行为，必须改变以业务经验驱动为核心的监测模式，尝试从数据出发，通过深入分析，挖掘出隐藏在大量正常数据中的异常查询行为。

2 以业务经验驱动的异常查询监测

根据业务经验，结合采集数据之间的逻辑关系，个人征信系统曾总结5条异常查询监测规则。

（1）查询量波动阈值

根据查询网点统计最近3年的日均查询量增长率，预设来年日查询量最大值，一旦超出阈值，视为异常。

（2）睡眠用户异常查询

在最近一年内均未发生过查询行为的用户，一旦启动查询操作，定义为睡眠用户异常查询。

（3）非工作时段异常查询

在过去非工作时段曾经发生查询行为的用户，若继续在非工作时段查询个人信用报告，视为异常。

（4）未授权异常查询

在未取得信息主体授权的情况下，用户以贷后管理为由查询非本行老客户的个人信用报告，定义为未授权异常查询。

（5）跨地域异常查询

商业银行以“贷款审批”为由查询个人信用报告，但用户所属机构的清算代码（金融机构代码第6、第7位）归属地与信息主体身份证号码归属地不是同一个省（自治区、直辖市），标记为异常；中国人民银行临柜用户以“本人查询”、“异议查询”为由查询个人信用报告，但用户所属机构的行政区划代码（机构代码第7、第8位）与信息主体身份证号码归属地不是同一个省（自治区、直辖市），标记为异常。此规则暂不考虑所在地为北京、上海、天津和广东等外来人口占比较大的区域内的机构。

通过以上5条监测规则，每月可侦测到上千万条疑似异常查询行为，并通过派出机构进一步核实，但反馈结果出乎意料，被核查的用户均给出合理解释，出现低检测率和高误报率的现象，使得监测行为陷入被动状态。

经过分析，其主要原因有以下两个方面。一是征信环境不断变化且各地发展不均衡，导致业务经验与实际情况存在较大

的时滞，未能准确地反映目前的情况。例如：部分商业银行以家庭为单位进行综合授信，放款前既要查询贷款本人的个人信用报告，又要查询贷款人家庭成员的个人信用报告，导致上文中提到未授权查询规则不适合实际业务情况；而与此同时，随着流动人口比重不断增加，跨地域查询也是合情合理的需求。二是没有深度数据分析支持的结论不具有说服力，没有针对性，难以被用户接受。例如：以“一刀切”的方式预设一个查询峰值，常常与实际查询需求相冲突，使该条规则饱受诟病。

3 以数据挖掘驱动的异常查询实时监测模型

根据数据挖掘算法，异常查询实时监测模型包含6个步骤：业务理解、质量检查、数据准备、数据分析、模型建立、模型验证。

3.1 业务理解

通过业务调研、违规查询样本分析及数据探索后发现，大量违规行为伴随查询量突增。典型案例如下：2015年3月某银行违规查询了3.2万份个人信用报告。从该用户的历史查询轨迹来看，原本平稳的查询频率在事发月份出现了异常突增，足以引起高度关注。类似的情况在多个案例中反复出现。因此，本次数据挖掘的目标定位为对用户月查询量进行预测，通过比对预测查询量与实际查询量的差异，判别用户的异常风险。

3.2 质量检查

检查查询记录的各字段值是否符合业务逻辑，并清理脏数据，保证后续的数据分析得出可靠的结论。

3.3 数据准备

本次数据挖掘的样本选用个人征信系统最新36个月全部查询记录，样本数为9.0亿条。经过数据预处理后，按月统计每个用户的查询总量，并形成查询量矩阵R。

$$R=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \quad (1)$$

查询量矩阵R共有m个月度观察点，n个查询用户，其中，$a_{ij}(1\leqslant i\leqslant n,1\leqslant j\leqslant m)$表示第$i$个用户在第$j$个月份的查询量。

3.4 数据分析

月查询量矩阵是一个稀疏矩阵，矩阵内存在大量为0的值，表明只有少量用户连续每个月都有查询，而大量用户的查询是时断时续的，因此有必要对查询连续性进行进一步的分析。

（1）查询休眠时长分析

分析用户最后一次查询距离当前日期的天数，定义为当前休眠天数T，统计结果显示：average(T)=197天，min(T)=0天，max(T)=973天。

结合**表1**与**图1**可以看出，50%和65%为突变点，$T\leqslant 84$天的用户达到50%，$T\leqslant 369$天的用户达65%。也就是说，最近3个月内，50%的用户至少发生过一次查

表1 不同休眠天数用户数占比

休眠天数T	0	2	84	369	973
用户数占比	0	25%	50%	65%	100%

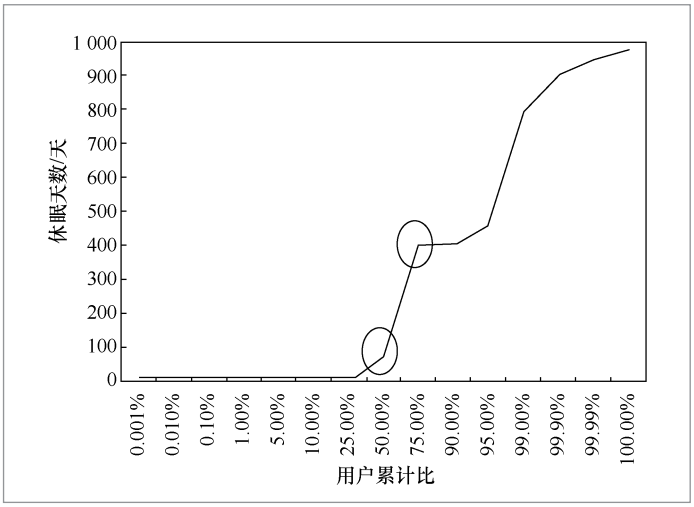


图1 用户休眠天数频度分析

询；最近12个月内，65%的用户至少发生过一次查询。

（2）休眠重启行为分析

用户休眠后又重新查询的行为特征是什么呢？以月为单位来计量，用户在自然月内有查询行为即为当月活跃，否则为休眠。滚动一个月后继续观察用户的活跃状态。由此分析正常用户的休眠、重启特征。

以2015年9月份的数据为例，当前活跃用户6.7万，占比42%。在2015年10月份，上个月6.7万活跃用户中，93%的用户继续活跃，剩余7%变成休眠1个月用户。2015年9月份休眠1个月用户0.4万，占比3%。在2015年10月份，这0.4万用户中，42%的用户又有了查询，再次活跃起来，剩余58%的用户由休眠1个月用户变成休眠2个月用户，具体见**表2**。

随着休眠时间增长，重启查询的可能性越来越低。为了验证结果的稳定性，依次对每个月的数据进行滚动分析，见**表3**和**图2**。

由**图2**可以看出，3个月、6个月也是与查询行为高度相关的特征值。休眠3个月的

表2 用户活跃数据

9月份状态		当前活跃	休眠1个月	休眠2个月	休眠3个月	休眠4个月	休眠5个月	休眠6个月	休眠7个月	休眠8个月	休眠9个月	休眠10个月	休眠11个月	休眠12个月
9月份用户数/万		6.7	0.4	0.3	0.4	0.8	0.4	0.2	0.1	0.1	0.2	0.1	0.1	6.1
9月份占比		42%	2%	2%	3%	5%	3%	1%	1%	1%	1%	1%	0	38%
10月份状态	再次活跃占比	93%	42%	18%	5%	2%	3%	4%	4%	3%	1%	3%	2%	4%
	休眠1个月	7%												
	休眠2个月		58%											
	休眠3个月			82%										
	休眠4个月				95%									
	休眠5个月					98%								
	休眠6个月						98%							
	休眠7个月							96%						
	休眠8个月								96%					
	休眠9个月									97%				
	休眠10个月										99%			
	休眠11个月											97%		
	休眠12个月												98%	
	休眠>12个月													96%

表 3 活跃率滚动数据

时间	当前活跃	休眠1个月	休眠2个月	休眠3个月	休眠4个月	休眠5个月	休眠6个月	休眠7个月	休眠8个月	休眠9个月	休眠10个月	休眠11个月	休眠12个月
2015年9月	93%	42%	18%	5%	2%	3%	4%	4%	3%	1%	3%	2%	4%
2015年8月	94%	38%	9%	3%	4%	6%	6%	5%	2%	2%	3%	3%	5%
2015年7月	93%	22%	6%	5%	7%	5%	5%	3%	4%	5%	3%	2%	5%
2015年6月	91%	17%	9%	10%	8%	9%	3%	5%	4%	3%	4%	0	6%
2015年5月	85%	20%	17%	10%	9%	4%	5%	6%	5%	4%	0	1%	6%
2015年4月	91%	38%	22%	20%	8%	9%	10%	5%	4%	0	1%	1%	8%
2015年3月	94%	41%	30%	12%	13%	12%	9%	6%	0%	1%	1%	2%	8%
2015年2月	96%	65%	24%	22%	18%	14%	8%	0%	2%	2%	2%	2%	9%
2015年1月	89%	25%	18%	12%	9%	5%	0%	1%	2%	2%	1%	1%	5%
均 值	92%	34%	17%	11%	9%	7%	6%	4%	3%	2%	2%	1%	6%

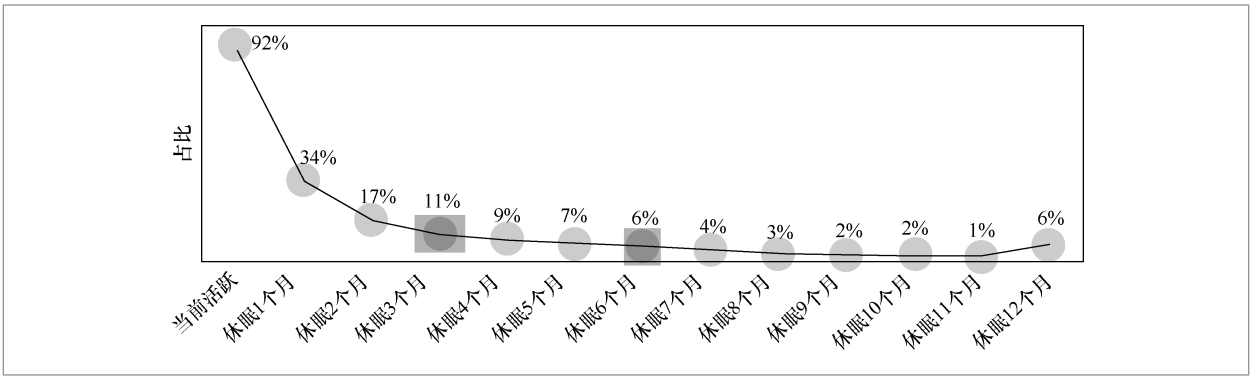


图 2 当前用户一个月后重启查询的平均占比

用户再次活跃的比例≤11%，休眠6个月的用户再次活跃的比例≤6%并趋于平稳。

3.5 模型建立

本次查询量预测目标需对每一个用户构建预测模型。

3.5.1 根据用户特征分组

查询矩阵中存在大量缺失值，建模前需对缺失值进行补充。为了能用最恰当的值补充，需要对用户按照查询特征进行分组。

根据上文用户查询特征数据分析结果，将用户活跃状态分为六大类，分别是活跃无断点、活跃有断点、新开用户、次新用户、当前休眠用户和历史休眠用户，具体见**表4**。

3.5.2 用户月查询量缺失值补充

以上六大类特征用户，其查询量缺失值补充规则见**表5**。

3.5.3 根据波动特征构建模型

原则上，用户的查询需求受其所属机构查询需求的影响，并保持相同趋势。用户所属机构分为十二大类，分别为：中国人民银行、全国性商业银行、城市商业银行、农村信用合作社、城市信用合作社、村镇银行、政策性银行、公积金管理中心、财务公司、汽车金融公司、外资银行、小额贷款公司。

表 4 用户活跃状态分类

序号	类型	特征
1	活跃无断点	观察期内均有查询记录
2	活跃有断点	观察期内存在无查询记录的月份，但连续不超过3个月
3	新开用户	只有最近6个月的查询记录
4	次新用户	有超过6个月但小于两年的查询记录
5	当前休眠用户	观察期的最近3个月连续无查询记录
6	历史休眠用户	其他

表 5 补充缺失值的规则

序号	类型	补充缺失值
1	活跃无断点	无需处理
2	活跃有断点	用相邻月份的均值填充
3	新开用户	待查询记录满6个月后再开始监测
4	次新用户	截取最新6个月的数据建模
5	当前休眠用户	直接设定阈值为200次，此类用户再次启用即需关注
6	历史休眠用户	用24个月非0查询量的均值填充

十二大类的机构呈现出4种不同的波动趋势，分别为平稳型、增长型、小幅跳跃型和周期跳跃型。因此，对用户的建模算法也遵循以上4类，其对应的预测算法见**表6**。

3.5.4 预测用户月查询量

用修正过的用户查询数据，根据用户所属组别选择预测模型，预测其月度查询峰值。

（1）平稳型

此类机构查询量每月基本稳定，使用简单平均算法来预测下个月的查询量，其计算式为：

$$\hat{X}_{24}=\frac{\sum_{i=0}^{23}X_i}{24}+N\times Std \qquad (2)$$

其中，$X_i(i=0,\cdots,23)$表示用户最近24个月的实际查询量；$\hat{X}_{24}$表示下一个月的月度预测值；N为调整系数；Std为24个月查询量的标准差。

（2）增长型

此类机构每月增长趋势明显，采用月度差分加权移动平均算法，其计算式为：

$$\hat{X}_{24}=X_{23}+\frac{\sum_{i=0}^{22}\Delta_i\times R^i}{\sum_{i=0}^{22}R^i}+N\times Std \qquad (3)$$

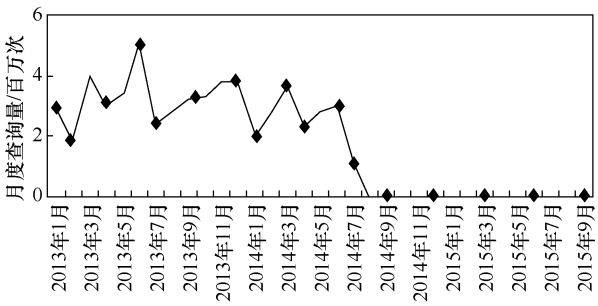
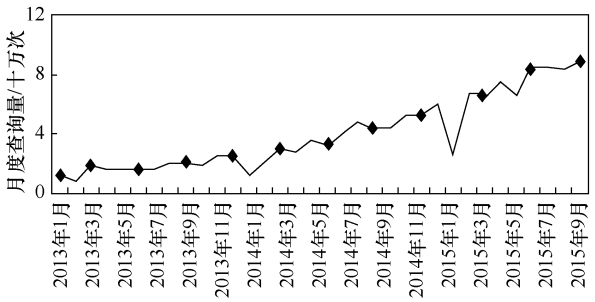
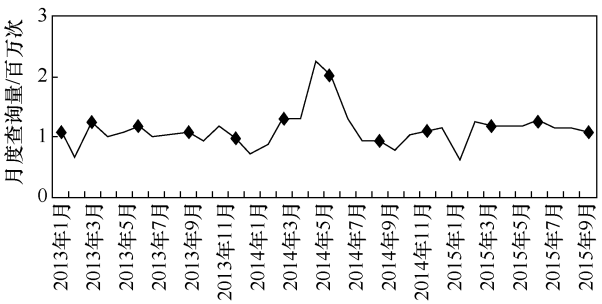
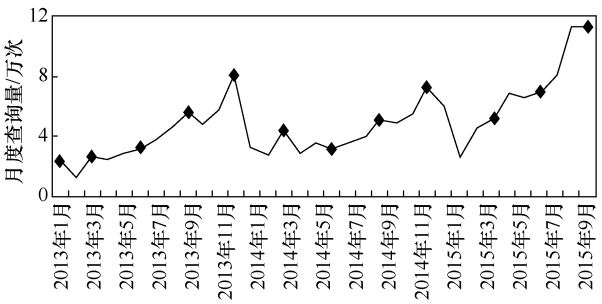
其中，$X_i(i=0,\cdots,23)$表示用户最近24个月实际查询量；$\hat{X}_{24}$表示下一个月的月度预测值；$\Delta i(i=0,\cdots,22)$表示每个月与上个月查询量的差值；N为调整系数；Std为24个月查询量的标准差；R表示指数权重，取值范围为(0,1)。

指数权重R的取值规则为：

$$W=\sum_{i=1}^{m}\left(X_{i24}-X_{i23}-\frac{\sum_{j=0}^{22}\Delta_{ij}\times r^i}{\sum_{j=0}^{22}r^i}\right)^2 \qquad (4)$$

其中，m为增长型用户总数；X_{i24}表示第i个用户预测值；X_{ij}表示第i个用户第j个月实际查询值；$r=0.01,0.02,0.03,\cdots,0.99$，每次

表 6 查询量波动特征及对应预测算法

类别	波动特征	算法
平稳型(全国性商业银行)		简单平均法 备注: 查询量锐减系当月开通接口查询
增长型(中国人民银行)		月度差分加权移动平均算法
小幅跳跃型(城市商业银行)		加权移动平均
周期跳跃型(财务公司)		年度差分算法

共计算99次; R的值是min(W)对应的r。

(3)小幅跳跃型

此类机构每月的查询量有一定的波动,但变化幅度很小,采用指数加权移动平均算法,其计算式为:

$$\hat{X}_{24}=\frac{\sum_{i=0}^{23}X_i\times R^i}{\sum_{i=0}^{23}R^i}+N\times Std \qquad (5)$$

其中, $X_i(i=0,\cdots,23)$表示用户最近24个月实际查询量; $\hat{X}_{24}$ 表示下一个月的月

度预测值；N为调整系数；Std为24个月查询量的标准差；R表示指数权重，取值范围为(0，1)。

指数权重R的取值规则为：

$$W=\sum_{i=1}^{m}\left(X_{i24}-\frac{\sum_{j=0}^{23}X_{ij}\times r^{j}}{\sum_{j=0}^{23}r^{j}}\right)^{2} \quad (6)$$

其中，m为小幅跳跃型用户总数；$\hat{X}_{i24}$表示第i个用户预测值；X_{ij}表示第i个用户第j个月的实际查询值；r=0.01,0.02,0.03，…,0.99，每次共计算99次；R的值是min(W)所对应的r。

(4)周期跳跃型

此类机构查询量波动有很强的周期性特征，采用年度差分方法，其计算式为：

$$\hat{X}_{36}=X_{24}+\sqrt{\frac{(X_{24}-X_{12})^{2}+(X_{12}-X_{0})^{2}}{2}}+N\times Std(X_{0},X_{1},\cdots,X_{35}) \quad (7)$$

其中，X_i(i=0,…,35)表示用户最近36个月实际查询量；$\hat{X}_{36}$表示下一个月的月度预测值；N为调整系数；Std为36个月实际查询量的标准差。

3.5.5 预测用户日查询峰值

月查询量预测即模型的结果，但为了满足实时监测的需要，需要将月度预测值推算至日预测峰值，具体计算式为：

$$\hat{d}_{24}=\frac{\sum_{i=0}^{23}d_i}{\sum_{i=0}^{23}X_i}\times\hat{X}_{24} \quad (8)$$

其中，X_i(i=0,…,23)表示用户最近24个月的实际查询量；$\hat{X}_{24}$表示下一个月的月度预测值；d_i为每个月的日查询峰值；$\hat{d}_{24}$表示下一个月预测的日查询峰值。

3.6 模型验证

利用相对误差及泰勒不等系数，对4种模型进行验证，具体见**表7**。结果表明，该模型应用于个人查询量预测是可行的。如平稳型模型相对误差的最大值、次大值和最小值分别为23.71%、23.47%、0.65%，平均精度为84.45%，泰勒不等系数为0.085，模型效果很好。

泰勒不等系数计算式为：

$$\frac{\sqrt{\sum_{i=1}^{n}(\hat{X}_i-X_i)^{2}/n}}{\sqrt{\sum_{i=1}^{n}\hat{X}_i^{2}/n}+\sqrt{\sum_{i=1}^{n}X_i^{2}/n}} \quad (9)$$

其中，n为预测期数，$\hat{X}_i$为预测值，X_i为实际值。

泰勒不等系数的值在0和1之间，当泰勒不等系数等于0时，是最优拟合。

平均相对误差计算式为：

$$\sum_{i=1}^{n}\left|\frac{\hat{X}_i-X_i}{X_i}\right|/n \quad (10)$$

其中，n为预测期数，$\hat{X}_i$为预测值，X_i为实际值。

4 模型应用及讨论

该模型针对每个用户可以得到2个预测值：一个是月度预测值$\hat{X}_{24}$；另一个是日

表7 模型验证结果

	最大相对误差	次大相对误差	最小相对误差	平均精度	泰勒不等系数
平稳型	23.71%	23.47%	0.65%	84.45%	0.085
增长型	45.52%	25.49%	25.02%	68.21%	0.213
小幅跳跃型	66.20%	34.63%	2.96%	73.86%	0.128
周期跳跃型	31.30%	28.08%	0.57%	85.09%	0.139

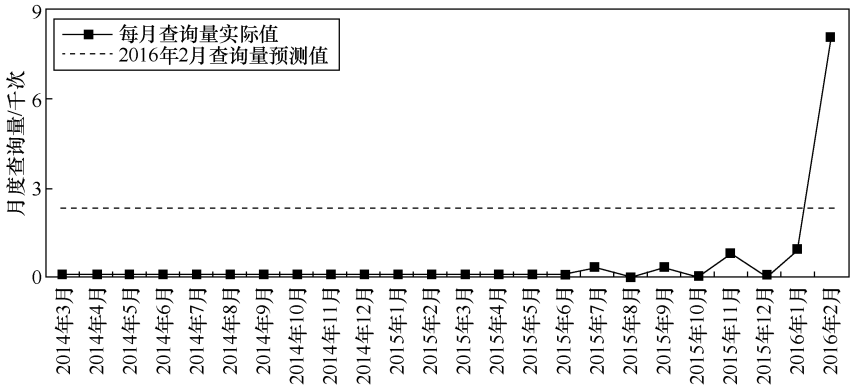


图 3 某用户最近 24 个月查询量趋势

查询峰值 $\hat{d}_{24}$。

为了实现实时监测的目标，系统每隔5 min从生产环境中提取查询记录，每次取数耗时4 s。然后按用户分别统计当月实际累计查询量Y和当日实际累计查询量e。监测结果W的计算式为：

$$W=\begin{cases}\text{true}, & (Y\geqslant\hat{X}_{24}\text{或者}e\geqslant\hat{d}_{24})\\ \text{false}, & (Y<\hat{X}_{24}\text{并且}e<\hat{d}_{24})\end{cases} \quad (11)$$

一旦W=true，系统自动告警，表示查询异常，并立马阻断该用户查询操作。

该数据挖掘模型部署至个人征信系统，从上线两个月的监测结果来看，共发现1 182笔异常查询行为。经过业务核查，确认存在违规查询。案例如下：该模型监测发现2016年2月份，XX中心支行某用户当月查询预测值为2 350次，而实际查询量为8 563次，查询异常，经核实确系违规操作，如图3所示。

5 结束语

本文结合前期业务驱动的监测结果及已掌握的异常查询案例，通过数据挖掘技术，从海量查询记录中，分析查询用户的行为模式，并归纳出平稳型、增长型、小幅跳跃型、周期跳跃型4种异常查询实时监测模型。经检验，该模型能快速准确地定位异常查询行为，从而更好地保护信息安全，同时，该模型成功上线后，对个人征信系统的违规查询行为产生威慑作用，倒逼查询机构加强内部管理，合法使用信用信息，以保障信息主体的权益，促进征信市场健康发展。

致谢

中国人民银行征信中心数据部高健、邓林慧、李状君、徐方林等同事对本研究工作给予了大量帮助，特此感谢。

参考文献:

[1] HAN J, KAMBER M. Data mining concepts and techniques[M]. Translated by FAN M, MENG X F. Beijing: China Machine Press, 2012.

[2] PANG-NING T, STEINBACH M, KUMAR V. Introduction to data mining[M]. Translated by FAN M, FAN H J. Beijing: Posts & Telecom Press, 2011.

[3] 中华人民共和国国务院. 征信业管理条例[M].

北京: 中国法制出版社, 2013.
The State Council of the People's Republic of China. Credit reporting industry regulations[M]. Beijing: China Legal Publishing House, 2013.

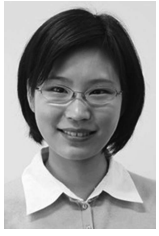
[4] 汪路. 论征信的本质及其主要特征[J]. 西部金融, 2010(6): 60-62.
WANG L. On the essence and main features of credit reference [J]. West China Finance, 2010(6): 60-62.

作者简介

姚前(1970-),男,中国人民银行征信中心副主任、高级工程师,主要研究方向为分布式系统和计算机安全。

谢华美(1976-),男,中国人民银行征信中心数据部副总经理,主要研究方向为数据挖掘。

景志刚(1977-),男,现就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。



胡青青(1984-),女,现就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。

司恩哲(1985-),男,现就职于中国人民银行征信中心数据部,主要研究方向为数据挖掘。

收稿日期: 2016-02-17