(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 106649527 A
(43)申请公布日 2017.05.10

(21)申请号 201610915505.X

(22)申请日 2016.10.20

(71)申请人 重庆邮电大学
地址 400065 重庆市南岸区南山街道崇文路2号

(72)发明人 刘群 谭敢锋 戴大祥

(74)专利代理机构 重庆市恒信知识产权代理有限公司 50102
代理人 刘小红

(51)Int.Cl.
G06F 17/30(2006.01)
G06Q 30/02(2012.01)

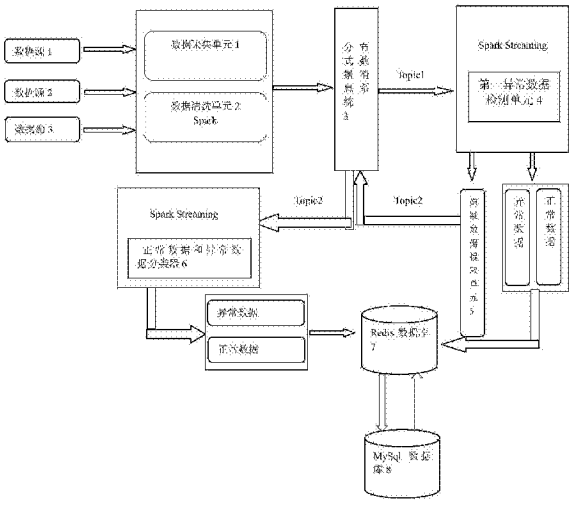
权利要求书2页 说明书4页 附图3页

(54)发明名称

基于Spark Streaming的广告点击异常检测系统及检测方法

(57)摘要

本发明请求保护一种基于Spark Streaming的广告点击异常检测系统及检测方法，涉及计算机技术应用领域，在用户点击网站广告时进行日志收集，对实时收集的数据进行清洗，标准化数据字段格式，然后将标准化数据由Flume传输给Kafka数据消息系统，Spark Streaming通过KNN邻近算法对数据进行分类，可以得到三大类数据异常数据、嫌疑数据、正常数据。对于异常数据和正常数据存储于数据库中，嫌疑数据发送给Kafka数据消息系统，然后通过异常数据训练朴素贝叶斯分类器，使用分类器可得到嫌疑数据的分类情况，数据保存于数据库中。最后，通过正常数据量合理收取广告商费用，同时可以分析得到各个广告的热门度，给广告商提供行业发展方向，提供用户全国分布情况等信息。



CN 106649527 A

1.一种基于Spark Streaming的广告点击异常检测系统，其特征在于，包括数据采集单元(1)、数据清洗单元(2)、分布式数据消息系统(3)、第一异常数据检测单元(4)、嫌疑数据提取单元(5)、正常数据和异常数据分类器(6)以及分类数据数据库单元；其中

数据采集单元(1)，用于采集用户点击广告的日志信息；

数据清洗单元(2)，对数据采集单元(1)采集到的日志进行清洗及标准化处理，最后将标准化后的数据发送到分布式数据消息系统(3)中，等待被消费；

分布式数据消息系统(3)，主要存储数据标准后的数据，还存储嫌疑数据提取单元发送来的的嫌疑数据，生成Spark Streaming所需消费的主题数据，不同的数据生成各自Topic；

第一异常数据检测单元(4)，采用了KNN算法对来自于分布式消息系统(3)中的数据在Spark Streaming中进行准实时处理，得到嫌疑数据、异常数据、正常数据；

嫌疑数据提取单元(5)，主要用于对第一异常数据检测单元(4)单元产生的嫌疑数据发送回分布式数据消息系统(3)中；

正常数据和异常数据分类器(6)，采用了朴素贝叶斯分类方法，对存储于分布式消息系统(3)的嫌疑数据进行分类，得到异常数据和正常数据；

分类数据数据库单元，包括括MySQL数据库(7)和Redis内存数据库(8)，其中MySQL数据库(7)用于存储正常数据和异常数据分类器(6)产生的正常数据和异常数据，并将异常数据映射给Redis内存数据库，便于快速训练朴素贝叶斯分类器，Redis为内存数据库，只是用于映射MySQL数据库，便于提高查询和修改的速度，设定一定周期内将数据写入到MySQL，便于永久保存。

2.根据权利要求1所述的基于Spark Streaming的广告点击异常检测系统，其特征在于，所述Redis内存数据库还包括将存储的异常数据用于进行训练朴素贝叶斯分类器。

3.根据权利要求1所述的基于Spark Streaming的广告点击异常检测系统，其特征在于，所述数据采集单元(1)采集用户点击广告的日志信息的设备为日志采集器Flume分布式日志收集系统，分布式数据消息系统为Kafka。

4.根据权利要求1所述的基于Spark Streaming的广告点击异常检测系统，其特征在于，所述第一异常数据检测单元(4)采用了KNN算法的KNN函数为：

$$y(x,c_j)=\sum_{d_i\in KNN} sim(x,d_i)y(d_i,c_j)-b_j$$

x为一条待分类日志的向量表示，d_i为训练集中的一条实例日志向量表示，c_j为一类别；它们的相似度使用余弦相似度，待分类日志和实例日志的相似度为：

$$\cos<x,d>=\frac{x\cdot d}{|x|\cdot|d|}$$

其中当d属于c_j时，取d为1，反之取0；距离度量使用欧几里得距离。

5.根据权利要求3所述的基于Spark Streaming的广告点击异常检测系统，其特征在于，KNN算法中，KNN分类器点击的有效性包括五个向量，第一个是“相同IP在一段时间内的点击数很多则异常”，第二个是“点击IP在广告页面的停留时间几乎可以忽略则异常”，第三个是“点击IP对于广告访问时刻异常的别于正常的人为活动时间”，第四个是“相同IP段不同地址访问同步性多次相似则异常”，第五是“对于IP行为和关注广告异常别于这个IP的以往行为和兴趣则嫌疑”，用这些样本数据作为KNN代表数据，得到KNN分类器。

6.根据权利要求3所述的基于Spark Streaming的广告点击异常检测系统，其特征在于，所述朴素贝叶斯函数为：

$$h_{nb}(x)=\arg\max_{c\in y}P(c)\prod_{i=1}^{d}P(x_i|c)$$

其中d为属性数目，x_i为x在第i个属性上的取值，

通过映射于Redis的异常数据为样本，训练该分类器，在一个周期内如：一周，就利用随机提取的20％的异常数据重新训练更新朴素贝叶斯分类器。

7.一种基于Spark Streaming的广告点击异常检测方法，其特征在于，包括以下步骤：

1)用分布式日志收集系统Flume采集网站用户的广告点击日志；

2)对步骤1)Flume采集到数据进行数据标准化处理，然后再由Flume将标准化数据发送到Kafka消息系统中，将这类原始的数据定义为Topic1，Topic1表示等待被消费数据，即相当于定义此类数据的地址；

3)对步骤2)中等待被消费数据Topic1，通过Spark Streaming准实时计算框架在KNN算法下进行分类；

4)根据步骤3)生成的嫌疑数据、异常数据、正常数据，将嫌疑数据发送回Kafka中定义为Topic2，其余数据保存于Redis内存数据库中，然后将这些数据写入MySQL数据库中，实现MySQL的读写分离；

5)根据步骤4)将Redis中随机提取于MySQL数据库中的20％的异常数据训练朴素贝叶斯分类器，然后将Kafka中的Topic2通过SparkStreaming准实时计算框架在朴素贝叶斯算法下进行分类。

8.根据权利要求7所述的基于Spark Streaming的广告点击异常检测方法，其特征在于，所述步骤3)中KNN算法为：将训练样本作为参考点，计算测试样本与训练样本的距离，采用欧氏距离，得到距离中最近的值作为分类的依据。

9.根据权利要求8所述的基于Spark Streaming的广告点击异常检测方法，其特征在于，步骤2)中所述KNN算法的欧氏距离的公式为：

$$dist(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

x和y表示差异个体，分别有n维特征。

基于Spark Streaming的广告点击异常检测系统及检测方法

技术领域

[0001] 本发明涉及计算机技术应用领域，具体是基于Spark Streaming广告点击异常检测系统及检测方法。

背景技术

[0002] 随着数据爆发式的增长，大数据的时代已来临，安全、快速、实时、高效的数据处理，不仅能够让企业提前规避风险，而且能够及时提供数据信息为企业发展，产品生产和开发提供真实有效的依据。

[0003] 然而，由于网络具有开放性，在方便大众的同时也带来了信息不真实、恶意访问、恶意攻击等。这是各个开放网站都面临的问题，怎样防止这些问题，怎样提取真实有效数据，减轻服务器恶意荷载是各个开放性网站的研究重点。其中投放广告的恶意点击就是一种典型问题，及时掌握异常数据阻止恶意点击，获得有效的广告点击数据，对开放性网站的合理收费提供依据，能够有效改善服务器负载，为投放广告商户提供合理的商业规划和业务指导具有重要意义。当下的处理技术，一般是基于离线批处理，这样的处理技术不能实时的解决线上问题，对某些需快速决策方案无法快速给出理论依据。对于实时型系统如：Storm，它虽然具备实时处理数据的能力，但是在数据安全性和大批量的数据处理上效果表现弱于Spark Streaming。Spark是一个类似于MapReduce的分布式计算框架，其核心是弹性分布式数据集，提供了比MapReduce更丰富的模型，可以在快速在内存中对数据集进行多次迭代，以支持复杂的数据挖掘算法和图形计算算法。Spark Streaming是一种构建在Spark上的实时计算框架，它扩展了Spark处理大规模流式数据的能力。

[0004] Spark Streaming的优势在于：

[0005] • 能运行在100+的结点上，并达到毫秒级延迟。

[0006] • 使用基于内存的Spark作为执行引擎，具有高效和容错的特性。

[0007] • 能集成Spark的批处理和交互查询。

[0008] • 为实现复杂的算法提供和批处理类似的简单接口。

[0009] 所以基于以上问题，结合现有的Spark大数据计算框架，及强大的电脑硬件支撑，合理的机器学习算法，能够快速、高效、精准的解决此类问题。

[0010] 本发明的一个目的就是提供基于Spark Streaming广告点击异常检测系统，它可以对投放于用户端的广告点击异常进行分析过滤，及时掌握有效广告点击情况，合理有效的广告投放计费，分析异常数据的行为和特征，更有助于分析用户行为和兴趣，为广告投放商提供商业规划，产品合理性等起到了事实依据，预测市场未来行情等。

发明内容

[0011] 本发明旨在解决以上现有技术的问题。提出了一种能够快速、高效、精准的为广告投放商提供商业规划、产品合理性等起到了事实依据、预测市场未来行情的基于Spark Streaming的广告点击异常检测系统及检测方法。本发明的技术方案如下：

[0012] 一种基于Spark Streaming的广告点击异常检测系统，其包括数据采集单元、数据清洗单元、分布式数据消息系统、第一异常数据检测单元、嫌疑数据提取单元、正常数据和异常数据分类器以及分类数据数据库单元；其中

[0013] 数据采集单元，用于采集用户点击广告的日志信息；

[0014] 数据清洗单元，对数据采集单元采集到的日志进行清洗及标准化处理，最后将标准化后的数据发送到分布式数据消息系统中，等待被消费；

[0015] 分布式数据消息系统，主要存储数据标准后的数据，还存储嫌疑数据提取单元发送来的的嫌疑数据，生成Spark Streaming所需消费的主题数据，不同的数据生成各自Topic；

[0016] 第一异常数据检测单元，采用了KNN算法对来自于分布式消息系统(3)中的数据在Spark Streaming中进行准实时处理，得到嫌疑数据、异常数据、正常数据；

[0017] 嫌疑数据提取单元，主要用于对第一异常数据检测单元单元产生的嫌疑数据发送回分布式数据消息系统中；

[0018] 正常数据和异常数据分类器，采用了朴素贝叶斯分类方法，对存储于分布式消息系统的嫌疑数据进行分类，得到异常数据和正常数据；

[0019] 分类数据数据库单元，包括括MySQL数据库和Redis内存数据库，其中MySQL数据库用于存储正常数据和异常数据分类器产生的正常数据和异常数据，并将异常数据映射给Redis内存数据库，便于快速训练朴素贝叶斯分类器，Redis为内存数据库，只是用于映射MySQL数据库，便于提高查询和修改的速度，设定一定周期内将数据写入到MySQL，便于永久保存。简而言之，Redis为一个中间件，为了提高速度而已。

[0020] 进一步的，所述Redis内存数据库还包括将存储的异常数据用于进行训练的朴素贝叶斯分类器。

[0021] 进一步的，所述数据采集单元采集用户点击广告的日志信息的设备为日志采集器Flume(分布式日志收集系统)，分布式数据消息系统为Kafka。

[0022] 进一步的，所述第一异常数据检测单元(4)采用了KNN算法的KNN函数为：

[0023] $$y(x,c_j)=\sum_{\vec{d}_i\in KNN} sim(x,d_i)y(d_i,c_j)-b_j$$

[0024] x为一条待分类日志的向量表示，d_i为训练集中的一条实例日志向量表示，c_j为一类别；它们的相似度使用余弦相似度，待分类日志和实例日志的相似度为：

[0025] $$\cos<x,d>=\frac{x\cdot d}{|x|\cdot|d|}$$

[0026] 进一步的，KNN算法中，KNN分类器点击的有效性包括五个向量，第一个是“相同IP在一段时间内的点击数很多则异常”，第二个是“点击IP在广告页面的停留时间几乎可以忽略则异常”，第三个是“点击IP对于广告访问时刻异常的别于正常的人为活动时间”，第四个是“相同IP段不同地址访问同步性多次相似则异常”，第五是“对于IP行为和关注广告异常别于这个IP的以往行为和兴趣则嫌疑”，对这些样本数据对KNN进行训练，得到KNN分类器。

[0027] 进一步的，所述朴素贝叶斯函数为：

[0028] $$h_{nb}(x)=\arg\max_{c\in y}P(c)\prod_{i=1}^{d}P(x_i|c)$$

[0029] 其中d为属性数目，x_i为x在第i个属性上的取值。

[0030] 通过映射于Redis的异常数据为样本，训练该分类器，在一个周期内如：一周，就利用随机提取的20％的异常数据重新训练更新朴素贝叶斯分类器。

[0031] 一种基于Spark Streaming的广告点击异常检测方法，其包括以下步骤：

[0032] 1)用Flume(分布式日志收集系统)采集网站用户的广告点击日志；

[0033] 2)对步骤1)Flume采集到数据进行数据标准化处理，然后再由Flume将标准化数据发送到Kafka消息系统中，将这类原始的数据定义为Topic1，Topic1表示等待被消费数据，即相当于定义此类数据的地址；

[0034] 3)对步骤2)中等待被消费数据Topic1，通过Spark Streaming准实时计算框架在KNN算法下进行分类；

[0035] 4)根据步骤3)生成的嫌疑数据、异常数据、正常数据，将嫌疑数据发送回Kafka中定义为Topic2，其余数据保存于Redis内存数据库中，然后将这些数据写入MySQL数据库中，实现MySQL的读写分离；

[0036] 5)根据步骤4)将Redis中随机提取于MySQL数据库中的20％的异常数据训练朴素贝叶斯分类器，然后将Kafka中的Topic2通过Spark Streaming准实时计算框架在朴素贝叶斯算法下进行分类。

[0037] 进一步的，所述步骤3)中KNN算法为：将训练样本作为参考点，计算测试样本与训练样本的距离，采用欧氏距离，得到距离中最近的值作为分类的依据。

[0038] 进一步的，步骤2)中所述KNN算法的欧氏距离的公式为：

[0039] $$dist(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

[0040] x和y表示差异个体，分别有n维特征。

[0041] 本发明的优点及有益效果如下：

[0042] 本发明通过Flume采集用户端投放广告点击数据，对数据进行清洗标准化，Flume对标准化后的数据发送到分布式消息系统Kafka中，等待订阅被消费生成Topic1，利用大数据准实时流数据Spark Streaming计算框架结合KNN分类算法，将数据分类为嫌疑数据、异常和正常数据，然后将嫌疑数据发送回Kafka中生成Topic2，同样利用大数据准实时流数据Spark Streaming计算框架结合朴素贝叶斯分类算法，将嫌疑数据生成的Topic2进行分类，得到异常数据和正常数据。在这些过程最终分类保存在Redis中，然后存储于MySQL数据库中，实现数据库的读写分离，增加读写速度。

附图说明

[0043] 图1是本发明提供优选实施例的结构示意图；

[0044] 图2为Spark Streaming下的KNN分类流程图；

[0045] 图3为Spark Streaming下的朴素贝叶斯分类流程图。

具体实施方式

[0046] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、详细地描述。所描述的实施例仅仅是本发明的一部分实施例。

[0047] 本发明的技术方案如下：

[0048] 如图1所示，一种基于Spark Streaming的广告点击异常检测系统，其特征在于，包括数据采集单元1、数据清洗单元2、分布式数据消息系统3、第一异常数据检测单元4、嫌疑数据提取单元5、正常数据和异常数据分类器6以及分类数据数据库单元；其中

[0049] 数据采集单元1，用于采集用户点击广告的日志信息；

[0050] 数据清洗单元2，对数据采集单元1采集到的日志进行清洗及标准化处理，最后将标准化后的数据发送到分布式数据消息系统3中，等待被消费；

[0051] 分布式数据消息系统3，主要存储数据标准后的数据，还存储嫌疑数据提取单元发送来的的嫌疑数据，生成Spark Streaming所需消费的主题数据，不同的数据生成各自Topic；

[0052] 第一异常数据检测单元4，采用了KNN算法对来自于分布式消息系统3中的数据在Spark Streaming中进行准实时处理，得到嫌疑数据、异常数据、正常数据；

[0053] 嫌疑数据提取单元5，主要用于对第一异常数据检测单元4单元产生的嫌疑数据发送回分布式数据消息系统3中；

[0054] 正常数据和异常数据分类器6，采用了朴素贝叶斯分类方法，对存储于分布式消息系统3的嫌疑数据进行分类，得到异常数据和正常数据；

[0055] 分类数据数据库单元，包括括MySQL数据库7和Redis内存数据库8，其中MySQL数据库7用于存储正常数据和异常数据分类器6产生的正常数据和异常数据，并将异常数据映射给Redis内存数据库，便于快速训练朴素贝叶斯分类器，Redis为内存数据库，只是用于映射MySQL数据库，便于提高查询和修改的速度，设定一定周期内将数据写入到MySQL，便于永久保存。简而言之，Redis为一个中间件，为了提高速度而已。

[0056] 图2为Spark Streaming下的KNN分类流程图。

[0057] 图3为Spark Streaming下的朴素贝叶斯分类流程图。

[0058] KNN分类器对标准化后存储于Kafka中的Topic1数据进行分类，生成嫌疑数据(KNN无法分类数据)，正常数据和异常数据，并将生成的正常数据和异常数据存储于数据库中，将嫌疑数据发送回Kafka中生成Topic2等待朴素贝叶斯分类器的分类，朴素贝叶斯分类器通过KNN分类后的异常数据进行训练，通过结合大数据Spark Streaming的超强计算能力，使计算变得更快，结果变得更精确，最后存储分类后的数据。

[0059] 本发明在网页用户点击广告后，实时过滤异常数据，并分析提取异常数据特征和行为，收集正常数据，合计计算广告投放费用，分析用户行为和兴趣，为广告投放企业制定商业策划，预测市场未来行情等。通过KNN的第一次分类达到三分类，嫌疑数据、异常数据和正常数据，然后通过异常数据对朴素贝叶斯进行训练，对嫌疑数据进行精确的划分，以达到数据的合理性，异常数据和正常数据，相关数据和非相关数据能够有力的为精确数据挖掘和预测分析提供保障。

[0060] 以上这些实施例应理解为仅用于说明本发明而不用于限制本发明的保护范围。在阅读了本发明的记载的内容之后，技术人员可以对本发明作各种改动或修改，这些等效变化和修饰同样落入本发明权利要求所限定的范围。

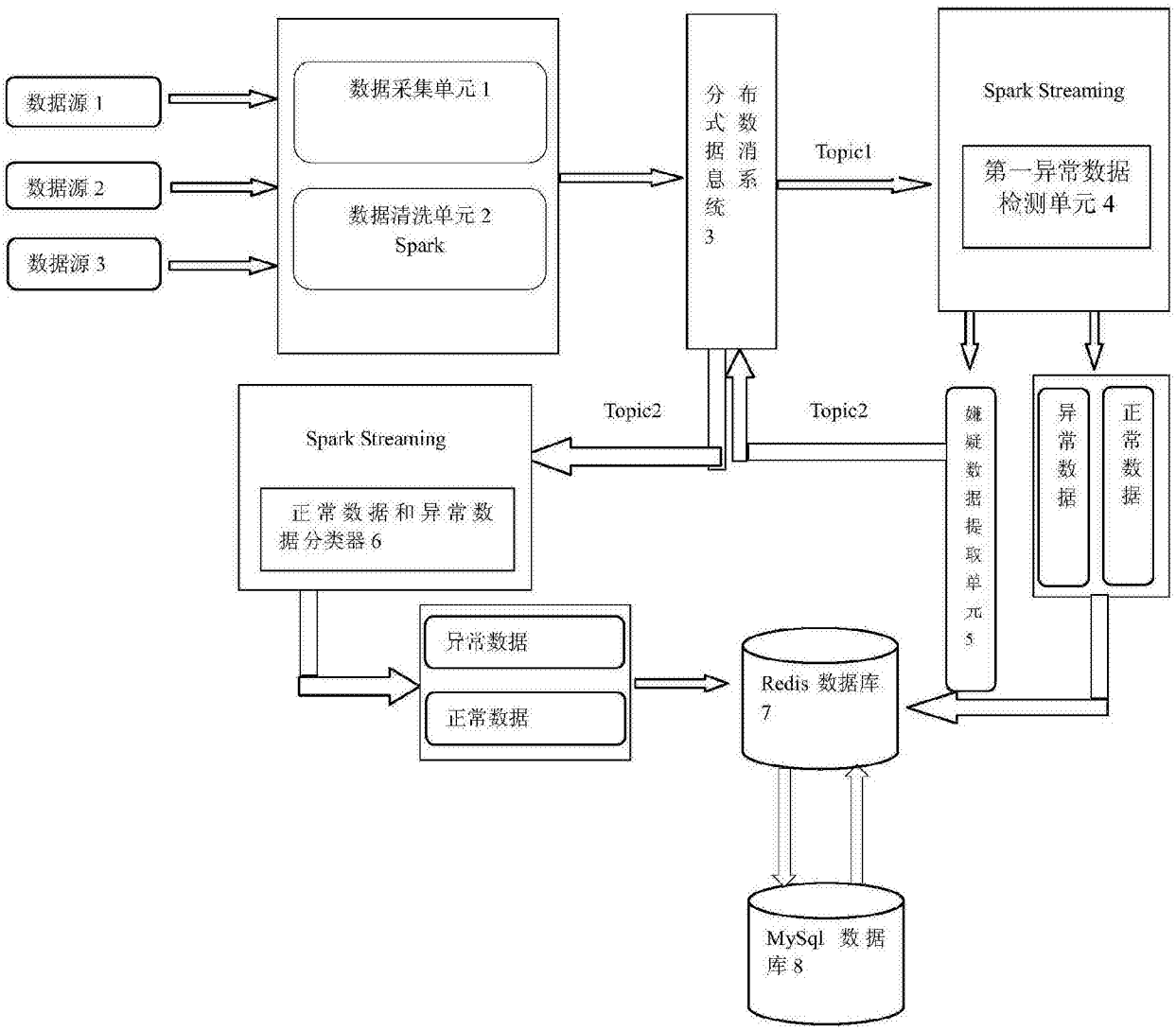

数据源 1
数据源 2
数据源 3
数据采集单元 1
数据清洗单元 2
Spark
分布式数据消息系统 3
Topic1
Spark Streaming
第一异常数据检测单元 4
Topic2
Topic2
Spark Streaming
正常数据和异常数据分类器 6
嫌疑数据提取单元 5
异常数据
正常数据
异常数据
正常数据
Redis 数据库 7
MySql 数据库 8

图1

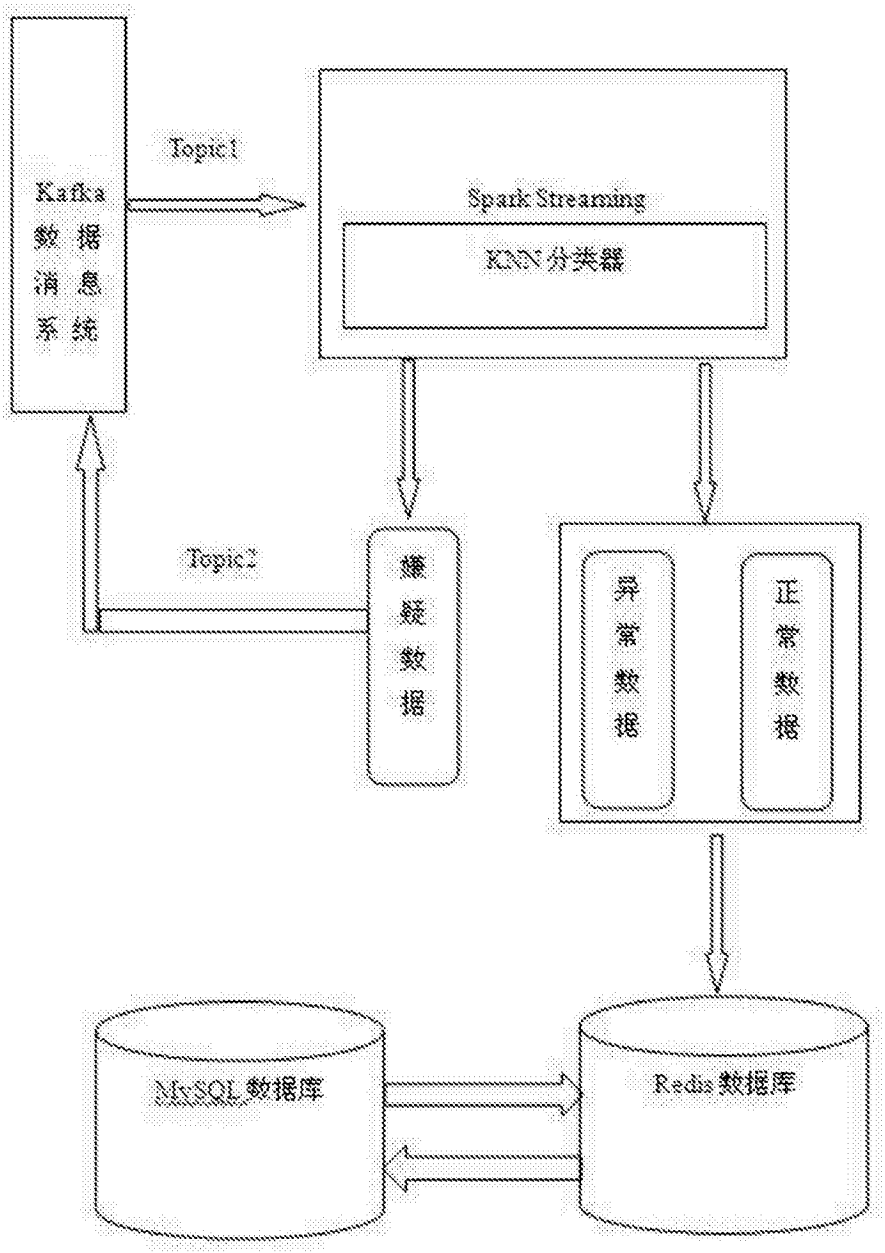

Kafka
数据
消息
系统
Topic1
Spark Streaming
KNN分类器
Topic2
嫌疑数据
异常数据
正常数据
MySQL数据库
Redis数据库

图2

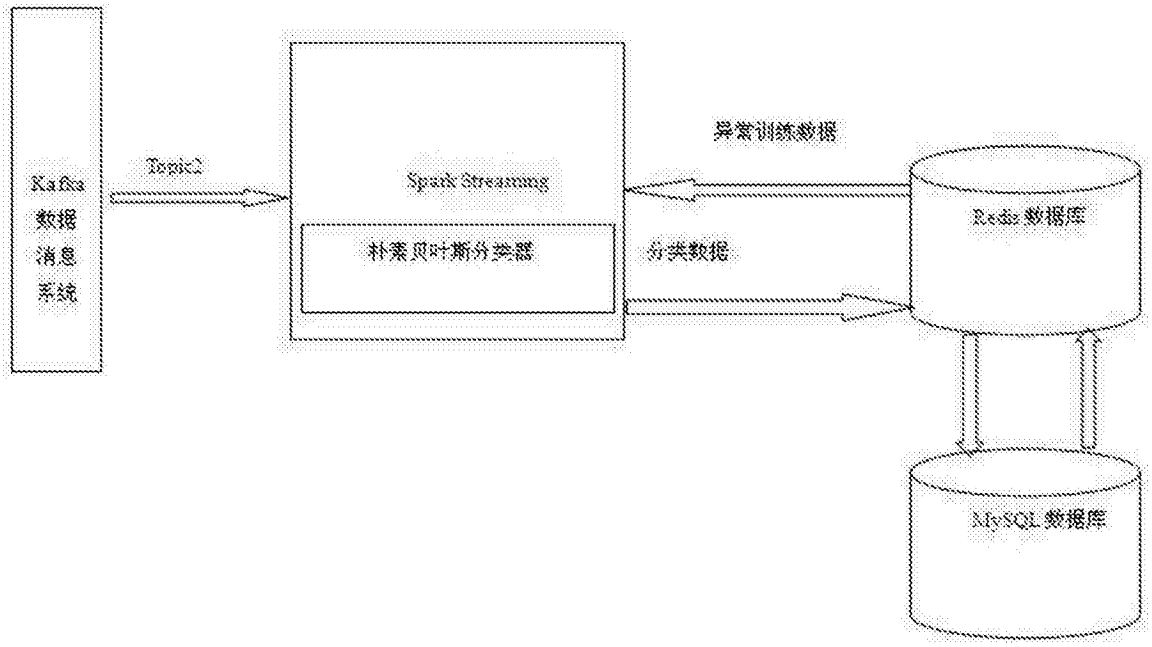

Kafka
数据
消息
系统
Topic2
Spark Streaming
朴素贝叶斯分类器
异常训练数据
分类数据
Redis 数据库
MySQL 数据库

图3