异常农产品价格数据的检测①

应　磊[1,2]　王儒敬[1]　杨紫微[1,2]　苏雅茹[1,2]

(1.中国科学院 合肥智能机械研究所 安徽 合肥 230031;2 中国科学技术大学 自动化系 安徽 合肥 230027)

摘　要： 农业垂直搜索引擎中的价格数据来源于各个农业网站，由于多种因素，采集到的数据中存在大量异常数据。同时，采集到的农产品价格数据具有其自身的特点，这些特点使得传统的异常数据检测方法不能够很好的工作。针对搜索引擎采集到的海量农产品价格数据，提出了一种异常价格数据的检测方法。实际应用表明，该方法能够取得很好的效果。

关键词： 异常数据；检测；垂直搜索；农业；价格数据

Detecting Abnormal Agriculture Price Data

YING Lei[1,2], WANG Ru-Jing[1], YANG Zi-Wei[1,2], SU Ya-Ru[1,2]

(1.Institute of Intelligent Machines, Chinese Academy of Science, Hefei 230031, China;

2. Department of Automation, University of Science and Technology of China, Hefei 20027, China)

Abstract: Agriculture price data from vertical search engines are collected from various agricultural sites. Due to various factors, there are a large number of abnormal data. Since the characteristics of these agriculture price data make traditional methods of detecting abnormal data not work well, this paper presents a new method of detecting abnormal data. Practical applications show that this method can achieve good results.

Keywords: abnormal data; detecting vertical search engines; agriculture; price data

1　引言

农业垂直搜索引擎就是要向用户提供专业的、可信的、精准的农业数据。随着互联网的快速发展，互联网上的信息呈爆炸式增长，在信息量增加的同时，一些瑕疵数据也在急剧增加，比如一些不完整数据，甚至一些错误数据也在大量增加[1]。因此，如何把搜索引擎采集到的数据进行适当的处理，检测出这些异常数据就显的尤为重要。

目前对于异常数据的检测方法，主要有：基于统计的方法、基于距离的方法、基于偏离的方法以及基于聚类的方法，但是这些方法都不适用于处理海量数据。在搜索引擎系统中，每天采集到的价格数据量超过 30000 条,并且所采集到的价格数据具有其自身特性。因此，本文提出了一种针对农业垂直搜索引擎采集到的价格数据的异常数据检测方法，该方法已经应用到系统中，并取得了显著效果。

2　问题描述与解决方案

2.1 问题描述

垂直搜索引擎采集到的农产品价格数据的主要属性有：农产品品种名称，农产品的交易地点(包括省、市、地区、农产品市场)，信息发布时间，发布信息的网站，农产品的价格，农产品价格的单位等。农产品价格数据具有其自身的特性，这些特性主要有：空间特性和时间特性。所谓的空间特性指：不同地区相同时间相同农产品的价格是有差异的。时间特性是指：

① 基金项目:国家高技术研究发展计划(863)(2006AA10Z23702,2006BAD10A0502);国家科技支撑计划(2006BAD10A1410);国家自然科学基金(60774096)

收稿时间:2009-08-12;收到修改稿时间:2009-09-07

同一个地区不同时间的相同农产品价格是不一样的。比如 A 地区和 B 地区同一时间某一种农产品的价格有一定差异；某地区今天某个农产品价格可能会跟几天前的价格有一定差异。

由于农产品的价格数据具有空间特性以及时间特性，因此我们处理数据时要充分考虑这两个特性。首先，我们应该按空间特性把采集来的海量数据进行区分，因为一条数据只有跟与它具有相同地区并且具有相同品种名称的数据进行比较才有意义。其次，我们处理时还应该充分考虑时间特性，对于农产品，同一个产品在夏季和冬季两个季节的价格会有很大变化。因此，我们需要把一条数据跟与它具有相同地区属性，相同品种名称属性以及具有相近时间属性的数据进行比较，才能够对其真实性进行判定。

由于农产品具有季节性，因此我们处理的数据集合会随着时间的变化而变化，比如西瓜的价格会随着季节的变化而逐步变化，因此，在处理数据的时候应该充分考虑这些数据的变化，即农产品价格的涨与跌。

对于农产品价格信息，还有一个数据集合上的特性：在某个时间段所采集的某个产品价格数据相对会集中在某一个数据点周围，这就是价格数据的相对集中性。因为某个地区某个产品的价格在一段时间内不会有很大的升降，而具有一定稳定性。

数据的相对集中性也给我们带来了处理上的问题。比如一周内北京地区黄瓜的价格可能会集中在 2.0 元/公斤。同样，在该段时间内，北京的青鳖的价格可能会集中在 70 元/公斤。在处理黄瓜价格数据信息的时候，如果采集到了一条数据是 8 元/公斤，那么我们有理由认为该条数据是有异常的。但是，对于青鳖的价格，假如一条数据是 80 元/公斤，虽然其与相对集中的数据差价达到了 10 元/公斤，我们是没有理由认定该条数据是异常的。因此我们需要一个不仅能够处理低价位农产品价格信息而且还能处理高价位农产品价格信息的方法。

农产品价格信息由于其具有空间特性，并且各个产品也有其自身特点，因此每次处理的价格信息的数量也是有变化的。比如 A 地区某个时间采集到的黄瓜的价格数据可能会达到 40 条，而采集到的金枪鱼的价格数据可能只有 4 条，因此我们也需要一个不仅能够处理小数量信息而且能够处理大数量信息的方法。

传统的异常数据检测方法在处理这些价格数据时，都不能够取得理想的效果，传统的方法在处理小数量信息时都不能够达到很好的效果，并且在每天采集到的数据中，小数量数据到达了 20%。其次，由于每天采集到的数据多达 30000 多条，因此我们还需要一个算法复杂度较低的方法，因此像基于聚类以及基于密度的异常数据监测方法也不能获得理想的效果。

2.2 解决方案

由以上问题描述可以知道，判断一条信息是否是异常信息，我们要把该条信息跟与它具有相同空间属性，相同品种属性，相似时间属性的数据进行比较，在此我们把相似时间属性定为最近一个周。因此，我们是把一个品种的价格数据跟该品种该地区前一个周的价格数据相对比，来判断其可信度。

数据的定义(以下数据定义都是某个地区某个品种某个时间段内的数据)：

Average 为最近一个周的平均价格(统计时应该去除这个周的最高最低价格，这样才能更好的进行判断)，Max 为最近一个周的最高价格，Min 为最近一个周的最低价格，Variance 为最近一个周的方差，Price 为当天我们要处理的一个价格，Max(a)表示 Max 是今天之前第 a 天的数据，当然 Max=Max(a)，在此只是为了描述算法的简便，t 表示当天的时间。

我们首先假定一周前的数据全是合理的，我们拿今天采集到的数据去跟之前一个周的数据进行比较，假如 Min<=Price<=Max，那么我们就认为该条数据是合理的，只有当 Price<Min 或者 Price>Max 时，我们才去进行进一步的判断。

在此，我们的判断方法有一个学习的过程(因为我们每天都要处理当天的数据)。假设 Max=Max(a)，如果今天的所有数据都在 Max 和 Min 之间，那么在明天处理该品种该地区的价格信息时，如果 a=7，则 Max 就有可能会减小而 Min 可能会增大，那么随着时间的推移，Max 和 Min 就会更为合理。

当 Price>Max，由于农产品价格数据在一段时间内是有可能涨的，故我们不能简单的断定其是否是异常数据。由于农产品数据具有相对集中性，那么 Average 就具有了很好的特性，它能很好的反映最近一段时间价格的分布聚集点，因此当一个数据距离 Average 很远的话，我们认为其很有可能是异常数据。

但是，由问题描述中的价格数据相对集中的特性知，我们不能简单的凭借距离就能判断其是否是异常数据，还要根据 Average 与 Min 数据的距离，因此，我们利用 p=(Price-Average)/(Average-Min)来判断。

但是，仅仅用 p 也不能判定其是否是异常数据，例如问题描述中所说的黄瓜和青鳖的数据。因此，我们还要根(Price-Average)/Average 的值来判定。

2.3 算法描述

根据前文描述，算法的流程如下：

(1) 如果数据库中还有数据，则从数据库中读取一条数据，读取其时间、空间、价格、单位属性值，转到(2)；

(2) 根据该条数据中的品种名称，交易地点，采集时间，从已经建立的索引中读取与该条数据拥有相同品种名称，相同交易地点的前一个周的统计数据：Average、Max、Min，转到(3)；

(3) 如果步骤(2)中读取的数据为空，即当前时间之前的一周内该地区都没有该品种的价格数据，则认为该条数据不是异常数据，转到(1)；否则转到(4)；

(4) 如果 Min<=Price<=Max，则认为该条数据不是异常数据，转到(1)；否则转到(5)；

(5) 如果 Price>Max,转到(6)；否则转到(8)；

(6) 如果 (Price-Average)/(Average-Min)<2.0,则认为该条数据不是异常数据，转到(1)；否则转到 7；

(7) 如果(Price-Average)/Average<0.5，则认为该条数据不是异常数据，转到(1)；否则认为该条数据是异常数据，转到(11)。

(8) 如果 Price<Min，转到(9)；

(9) 如果 (Average-Price)/(Max-Average)<2.0，则认为该条数据不是异常数据，转到(1)；否则转到 10；

(10) 如果(Average-Price)/Average<0.5，则认为该条数据不是异常数据，转到(1)；否则认为该条数据是异常数据，转到(11)。

(11) 对异常数据进行标记并转到(1)。

3 实验结果及分析

把该算法应用到农业垂直搜索引擎“搜农”(www.sounong.net)中，获得了良好的效果。“搜农”每天采集到的数据经过去重、消除不完整数据等前期处理后，每天的价格数据超过 30000 条。我们比较了本文提出的方法、基于正态分布[2]的方法以及基于偏离的方法[3]，我们通过比较该三种方法的错误判定率来比较他们的可靠性。

所谓的错误判定就是指本来不是异常的数据，却把它判定为异常数据，错误判定率 p=错误判定的数据量/判定为错误数据的数据量。对于以下数据，我们认为是错误判定的：

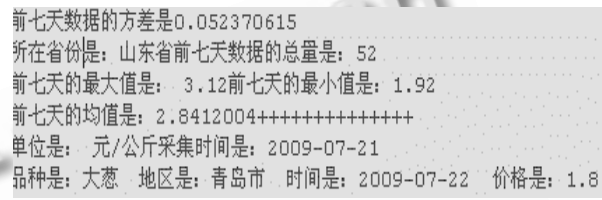

图 1　本算法在“搜农”系统中的运行结果

“搜农”是一个使用 JAVA 基于 LUCENE 开发的农业垂直搜索引擎。对于这条数据，前七天的最大值是 3.12，最小值是 1.92，均值是 2.8412004，方差只有 0.052370615。对于此条数据，实际上我们是不能把该条数据判定为异常数据的。对于该条数据，由于方差过小，当用基于正态分布的方法时[2]，显然会误判。因此，当把诸如该类数据判定为异常数据时，我们认为进行了一次错误判定。

表 1　三种检测方法的比较

方法	测试数据量	检测出的异常数据	误判数目
1	35421	641	388
2	35421	507	254
3	35421	279	26

由表 1 我们可以看出：

基于正态分布的方法的错误判定率是最高的，错误判定的数量达到了 388，而本文提出的方法的错误判定的数量是最少的，只有 26 条。

因此，本文所述方法能够很好的工作，达到了我们所期望的目标。

4 总结

本文详细描述了农业垂直搜索引擎采集到的价格

(下转第 207 页)

(上接第 180 页)

数据所具有的特点，仔细探讨了在该海量数据中检测异常数据所面临的问题，针对这些问题本文提出了一种检测异常农产品价格数据的方法，通过该方法在农业垂直搜索引擎“搜农”(www.sounong.net)中的应用表明，该方法能够很好的满足实际应用的要求。我们下一步的工作是对采集来的数据进行进一步的处理，寻找这些异常数据的根源，把异常数据进一步分类，把一些可以修复的异常数据进行修复。

参考文献

1 谭 · 斯坦巴赫.数据挖掘导论.范明,范宏建,译.北京:人民邮电出版社, 2006. 21－23.

2 陈希儒.高等数理统计学.合肥:中国学学学技术大学出版社, 1999. 5－20.

3 Arning A, Aggarwal C, Raghavan P. A Linear Method for Deviation Detection in Large Databases. Proc. Int’l Conf. Knowledge Discovery and Data Mining (KDD’96), 1996. C164－169.