

Benford 法则在水文数据质量挖掘中的应用研究

余宇峰^{1,2}, 万定生^{1,2}

(1 河海大学 计算机与信息学院, 江苏 南京 210098;

2 水资源高效利用与工程安全国家工程研究中心, 江苏 南京 210098)

摘 要: 为识别和改进数据中存在的质量问题, 采用 Benford 法则进行数据质量挖掘分析, 该方法通过分析数字分布规律来检测数据的合理性, 达到控制数据质量的目的. 以水文数据中降水量数据为样本验证方法的有效性. 实验结果表明, 该方法能有效识别数据集中存在异常信息, 提高了水文数据的数据质量, 具有一定的应用前景.

关键词: 水文数据; 数据质量; 数据挖掘; Benford 法则

中图分类号: TP302.1

文献标识码: A

文章编号: 1000-7180(2011)08-0180-04

An Application Research of Bedford's Law in Hydrological Data Quality Mining

YU Yu-feng^{1,2}, WAN Ding-sheng^{1,2}

(1 College of Computer & Information, HoHai University, Nanjing 210098, China;

2 National Engineering Research Center of Water Resources Efficient
Utilization Engineering Safety, Nanjing 210098, China)

Abstract: In order to identify and improve the data quality problems, this paper adopts a data quality mining methods based on Benford's Law. This method detects the rationality of data through analyzing the distribution of the data to reach the goal of data quality control. Finally, we used precipitation data as a sample to verify the validity of this method. The results show that the method can effectively identify the abnormal data, improving the data quality, and has certain application prospect.

Key words: hydrological data; data quality; data mining; Benford's Law

1 引言

在信息化时代, 数据是企业生产和运行的基础, 其质量好坏直接影响着企业的生存和效益. 据 Data Flux 和 SAS 研究表明, 美国企业每年要花费将近 6 000 亿美元在数据质量问题上. 低质量的数据往往导致错误的决策, 因此, 越来越多的部门在进行信息系统建设时重视起数据质量问题.

我国水文部门自 1980 年代开始实施信息化以来, 已经建成了一系列的数据库管理系统及数据采集系统, 积累了大量水文水资源科学数据, 为防汛抗旱、水资源管理和保护发挥了重大作用. 但由于原始

水文数据量庞大, 数据处理自动化程度不高; 以及数据建设过程中的存在的地域差异, 导致水文数据质量较低, 直接影响了水文系统的科学化决策进程. 如何对水文数据进行质量控制成为水文信息化建设过程中值得探讨和必须解决的问题.

2 相关理论研究

2.1 数据质量

数据质量(Data quality, DQ)通常被定义成“适合使用”的数据, 即收集的数据能满足特定用户需求^[1-2]. 在文献[3]中, 数据质量定义成一个信息系统中的数据模式和数据实例间相一致程度, 并在此基

收稿日期: 2011-05-19; 修回日期: 2011-06-27

基金项目: 国家自然科学基金项目(51079040); 水利部“九四八”项目(201016)

基础上实现数据的精确性(accuracy)、完整性(completeness)、有效性(validity)、一致性(consistency)等。为了解决数据质量问题,研究人员提出了很多数据质量分析方法。

数据质量分析是指在数据集成时,发现并解决质量问题的过程。最初,研究人员主要是在领域知识指导下,结合元数据、统计学、排序等技术进行数据质量分析,如相似重复记录识别、缺失值预测与填补、可疑数据或奇异点检测^[4-6]等。但大多质量问题隐藏在数据模式之中,不易被这些方法发现。由于数据挖掘在发现隐藏模式上的突出特点,近来利用数据挖掘方法进行数据质量分析逐渐成为研究重点。

2.2 数据挖掘

数据挖掘(Data Mining, DM)也称数据库中的知识发现(Knowledge Discovery in Database)。DM 试图从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中获取新颖的、潜在有用的、可理解的信息,这些信息表现为概念、规则、模式等;可以帮助决策者分析历史数据及当前数据,从中发现隐藏的关系和模式,进而预测未来可能发生的行为。

数据挖掘可以发现关联规则、特征规则、分类规则、聚类规则、离群数据等典型知识。常用的数据挖掘技术主要有神经网络、遗传算法、决策树方法、邻近搜索、规则归纳、模糊逻辑技术、可视化技术、粗集方法等。

2.3 基于数据挖掘的数据质量分析

基于数据挖掘的数据质量分析,是指应用数据挖掘技术进行数据质量分析和改进,其目的是为了检测、量化、解释和校正大规模数据库中存在的 data 质量问题,也叫做数据质量挖掘(data quality mining, DQM)^[7]或数据审计(data auditing)^[8]。DQM 可以解决很多数据质量问题,如利用预测方法进行缺失值分析,利用聚类和分类方法进行奇异点检测,利用聚类、分类和关联分析进行相似重复记录匹配等;发现领域知识和元数据之外的隐藏模式,辅助建立规则,发现隐藏的质量问题;提高分析的客观性和准确性,增强方法的通用性。

目前基于数据挖掘的数据质量分析研究主要集中于隐藏规则的挖掘与异常发现、相似重复记录匹配、冗余字段分析与合并规则挖掘数据库结构挖掘等方面。

3 Benford 法则

数字分析法是根据数字分布规律来检测异常,

达到控制数据质量的目的。而以 Benford 法则为基础发展起来的数字分析法,由于其在过去 20 年中在各个领域中的成功应用,正逐渐受到业内人士的关注,值得在数据质量挖掘领域推广。

3.1 Benford 法则概述

研究表明,一堆从实际生活得出的数据中,首位数为小数字的可能性比大数字的可能性大,这就是著名的 Benford 法则,它可用于检查数据是否存在质量问题。美国国家标准与技术学会给出了 Benford 法则的定义:在不同种类的统计数字中,首位数字是数字 d_1 的概率是^[9]

$$P(D = d_1) = \lg(1 + 1/d_1) \quad d_1 = 1, \dots, 9 \quad (1)$$

其中,首位数字是指左边的第一位非零有效数字。如数字 123, 1.23, 0.123 的首位数字均为 1。

根据式(1),还可以推导出第二位数字为 d_2 、第三位数字为 d_3 的概率分别为

$$P(D_2 = d_2) = \sum_{d_1=1}^9 \lg[1 + 1/(d_1 * 10 + d_2)] \quad d_2 = 0, 1, \dots, 9 \quad (2)$$

$$P(D_3 = d_3) = \sum_{d_1=1}^9 \sum_{d_2=0}^9 \lg[1 + 1/(100 * d_1 + 10 * d_2 + d_3)] \quad d_3 = 0, 1, \dots, 9 \quad (3)$$

根据上述公式,可以直接计算出 0~9 出现的概率如表 1 所示。

表 1 基于 Benford 法则的数字分布概率

数字	首位数	第 2 位数	第 3 位数	第 4 位数
0		0.119 68	0.101 78	0.100 18
1	0.301 03	0.113 89	0.101 38	0.100 14
2	0.176 09	0.108 82	0.100 97	0.100 10
3	0.124 94	0.104 33	0.100 57	0.100 06
4	0.096 91	0.100 31	0.100 18	0.100 02
5	0.079 18	0.096 68	0.099 79	0.099 98
6	0.066 95	0.093 37	0.099 40	0.099 94
7	0.057 99	0.090 35	0.099 02	0.099 90
8	0.051 15	0.087 57	0.098 64	0.099 86
9	0.045 76	0.085 00	0.098 27	0.099 82

Benford 法则被提出以后,在大量的应用领域得到推广。但并非所有数据都遵循 Benford 法则, Nigrini^[10]提出用 Benford 法则进行分析的数据应该满足以下几个条件:数据量具备一定规模,能够代表所有样本,数据集规模越大, Benford 法则分析结

果越精确;数据不能有人为限定的最大值和最小值;数据受人为主的影响较小。

Nigrini 将该方法应用到税务、会计等领域,成功检测出大量舞弊案例。在统计领域,John Nye 和 Charles Moul 将该法则应用于经济合作与发展组织(OECD)成员国的 GDP 统计数据,检测出数据质量中存在问题。

3.2 Benford 法则的分析流程

Benford 法则为基础的数字分析法在数据质量挖掘过程中的一般流程:

- (1)对数据集进行分析,初步判断数据集是否符合 Benford 法则的分析条件,选择分析的目标字段;
- (2)对总体进行分析,进行首位有效数字 Benford 法则的符合性测试,初步发现测试线索和重点;
- (3)根据初步线索,进行首、二、三位数字的详细分析;
- (4)在需要的情况下,对数据进行细分,进行分组分析,按照(1)~(3)步骤深入挖掘;
- (5)将标准分布频率和实际分布频率进行比较,寻找异常偏差点;并分析出现的异常原因,解决数据质量问题。

4 水文数据质量挖掘应用

依据 Benford 法则理论,人为的误操作、测量错误等行为将破坏水文数据中的客观分布规律,使之出现异常。因此,将目标数据的数字分布规律与基于 Benford 法则的标准分布规律进行比较,然后运用非参数统计中的 χ^2 检验法估计每个数字出现的实际次数与期望次数之间分布的整体拟合程度,从而验证水文统计数据的前 i 位数字的分布规律与 Benford 法则下的期望规律是否相符。通过分析比较挖掘出水文数据中的异常点,从而达到控制水文数据质量的目的。

由于水位、流量和水质数据受人工干预因素较大,不适宜采用 Benford 法则进行挖掘分析。降雨量、蒸发量等水文要素受人工干预影响较小,可以应用该法则进行数据质量分析。

4.1 数据准备

选择太湖流域 1954—2006 年旬均降雨量数据作为数据挖掘分析的样本,原始数据记录共 76 860 条。为提高原始数据的精确度,保证分析过程的准确性,首先需要对取得的数据进行整理。

(1)按 Benford 法则数据分析法,剔除旬均降雨量数据为 0 的数据 8 160 条,得到有效记录

68 754 条。

(2)根据 Benford 法则“比例无偏性”的特点,为便于分析,将样本数据均扩大 10 倍转换成整数(原始数据精确到小数点后一位)。

4.2 基于 Benford 法则的数据质量挖掘

采用的样本数据 6 万余条,具备一定的规模,满足数据集规模的要求;降雨量数据没有最大值和最小值的限制,受人为影响也较小,符合 Benford 分布的规律。因此,可以利用 Benford 法则对该组数据进行分析。

4.2.1 首位有效数字的质量挖掘

按 Benford 法则的分析流程,对降雨量数据的首位数字出现的概率进行初步分析,将实际分布频率与标准分布频率进行比较,初步发现测试线索和重点。将所得样本数据的实际分布频率与标准分布频率进行比较,可以得到图 1 所示结果。

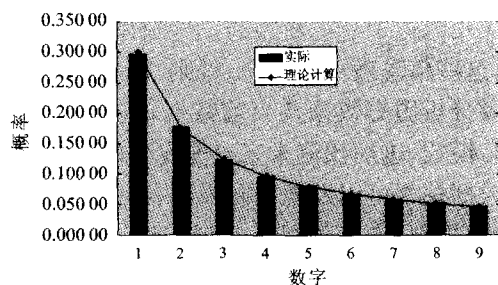


图 1 首位数字的 Benford 验证

由图 1 分析可以得出两点结论:

- (1)该样本数据记录在总体上符合 Benford 法则的分布规律;
- (2)在局部存在偏差,如在数字“1”、“3”、“7”处,差异显著。

4.2.2 前两位、前三位数字的详细分析

根据上述的分析,只能确定在“1”、“3”、“7”处存在显著差异,但是还难以直接定位异常数据点。可以通过深入挖掘,进一步查找异常值。

依次类推,对样本数据的前二位有效数字的组合进行分析,可以得到图 2 的结果。

从图 2 可以看出,数据总体的分布频率仍然比较符合 Benford 法则。在 10, 16, 39, 51, 66 等点处出现了一定的偏差,但仍然不具有显著性,可以进一步深入挖掘。对样本数据的前三位有效数字进行深入分析,得出的结果如图 3 所示。

从图 3 的曲线可以看出,在一些数据点存在较大的频率偏差,明显偏离 Benford 法则的分布规律。在这个基础上,可以进一步挖掘到详细的数据。将局部放大后如图 4 所示。

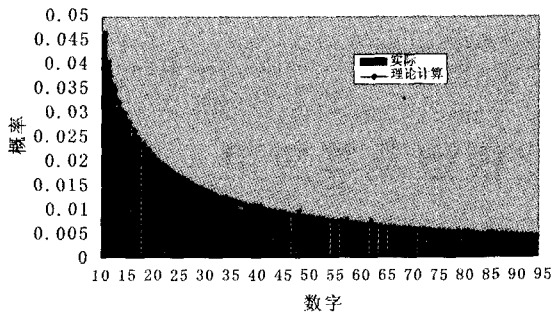


图2 首二位数字的 Benford 验证

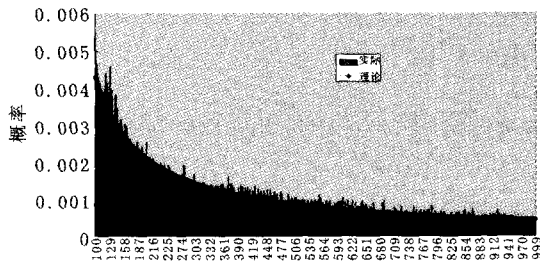


图3 首三位数字的 Benford 验证

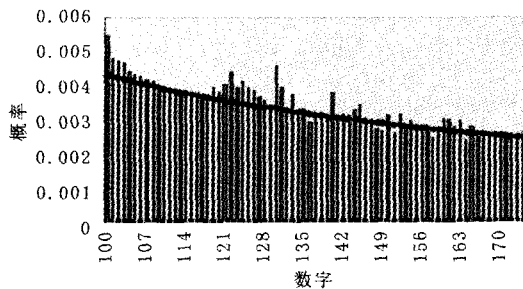


图4 局部放大图

如图4所示,前三位数为130的点与理论值存在较大的频率偏差,数据质量审核人员可以此为据对原始测量结果值为13.0的数据进行深入的排查和分析,解决原始数据记录可能存在的质量问题。

4.3 数据质量挖掘准确性评估

通过 Benford 进行数据质量挖掘后,能发现水文数据中存在的质问题。为对水文数据质量挖掘结果的准确性进行评估,

采用非参数统计中的 χ^2 拟合度检验来估计每个数字出现的实际次数与期望次数之间分布的整体拟合程度。 χ^2 拟合度检验公式如下:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

式中, O_i 和 E_i 分别是降雨量统计数据中数字 i 出现的实际次数和期望次数。 χ^2 统计值越大,说明降雨量统计数据越不符合 Benford 法则,其准确性就越值得怀疑。下面以首位数字的 χ^2 拟合度为例验证数据质量挖掘准确性评估方法。

H_0 : 旬均降雨量统计数据的首位数字实际出现

次数与 Benford 法则的期望出现次数没有显著差别;

H_1 : 旬均降雨量统计数据的首位数字实际出现次数与 Benford 法则的期望出现次数有显著差别。首位数字分布偏差情况如表2所示。

表2 首位数字 χ^2 偏差校验表

数字	O_i	E_i	$O_i - E_i$	χ^2
1	20 472	20 697	-225	2.452 594 2
2	12 204	12 107	97	0.778 230 8
3	8 439	8 590	-151	2.644 904
4	6 687	6 663	24	0.085 954 8
5	5 514	5 444	70	0.891 168 2
6	4 642	4 603	39	0.329 654 6
7	4 037	3 987	50	0.625 355 1
8	3 582	3 517	65	1.209 418 1
9	3 178	3 146	31	0.312 726 3
总计	68 754	68 754	0	9

在0.05的显著性水平下 $\chi_{0.005}^2(8) = 15.507$ 。由表2可知,首位数字分布的 χ^2 统计值9小于15.507。因此不能拒绝 H_0 ,即可以认为旬均降雨量统计数据的首位数字的分布规律与 Benford 法则期望规律相符。

5 结束语

本文将 Benford 法则引入到水文水资源数据的数据质量挖掘分析过程中,并以太湖流域1954—2006年旬均降雨量数据为例验证了该方法在水文数据质量挖掘中的效果。实验结果表明,该方法能有效挖掘水文水资源数据的质量问题,从而达到提高水文水资源数据质量的目标。但在使用时必须先判定待检数据是否满足 Benford 法则要求;若数据检验结果不符合 Benford 法则说明数据可能存在质量问题,可以此为线索进行数据质量分析,明确数据质量产生的原因,实现提高数据质量的目的。

参考文献:

- [1] Capiello C, Francalanci C, Pernici B. Data quality assessment from the user's perspective[C]//Proceedings of the 2004 international workshop on Information quality in information systems. New York, NY, USA: ACM, 2004.

(下转第186页)

75.6%, 82.3%; 图 2 中 GA, KDD 和 ICIRA 三者的平均查全率分别为 72.6%, 72.1%, 82.8%。两次对比实验均表明 ICIRA 检索算法的性能最好, 同时能获取比较高质量的检索文档。

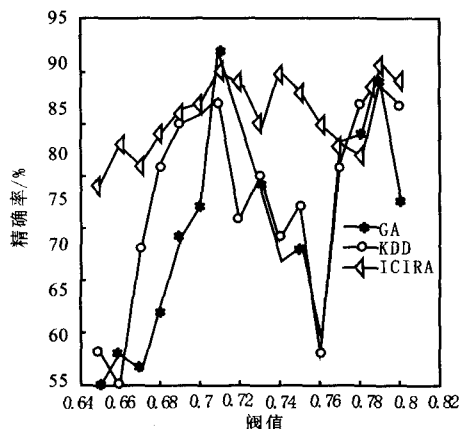


图 2 三种检索算法的查全率比较

5 结束语

随着 Internet 的飞速发展及其应用的日益普及, 网上的信息急剧膨胀, 要在这样一个浩如烟海的信息空间里查找所需的信息, 人们觉得力不从心, 往往花费了很多时间和精力却收获甚少。各种搜索引擎的出现是解决上述问题的主要途径。但是目前的搜索引擎检索出的信息量庞大, 常常混杂相当数量用户不感兴趣的, 甚至无关的信息。因此, 如何在 Internet 上获得高检索精确度和高检索率, 是当前 In-

ternet 资源发现的重要研究课题之一。本文将克隆的思想和方法引入 Web 搜索引擎, 能有效提高 Web 搜索引擎质量和效率, 具有很强的理论意义和实际价值。

参考文献:

- [1] 曾致远, 张莉. 基于向量空间模型的网页文本表示改进算法[J]. 计算机工程, 2006, 32(3): 134-136.
- [2] 凌波, 周水庚, 周傲英. P2P 信息检索系统的查询结果排序与合并策略[J]. 计算机学报, 2007, 30(3): 405-414.
- [3] 陈福集, 杨善林. 一种基于 KDD 的 Web 搜索引擎框架[J]. 情报学报, 2002, 21(3): 264-268.
- [4] 李新安, 石冰. 基于决策树方法的特定主题 Web 搜索策略[J]. 计算机应用, 2006, 23(1): 235-268.
- [5] 张卫丰, 徐宝文, 周晓宇. Web 搜索引擎综述[J]. 计算机科学, 2001, 28(9): 24-28.
- [6] 王晓峰, 刘惟一. 通过网页集团对 Web 信息搜索进行优化[J]. 云南大学学报: 自然科学版, 2003, 25(6A): 75-78.
- [7] Yang J R R. Effects of query term weights modification in document retrieval: A study base on a genetic algorithm[C]// proceedings of the second annual symposium on Document Analysis and Information Retrieval, New York, NY, USA: ACM, 1993: 271-185.

作者简介:

詹 芹 女, (1977-), 硕士, 讲师. 研究方向为数据挖掘、嵌入式系统。

(上接第 183 页)

- [2] Strong D, Lee Y, Wang R. Data quality in context[J]. Commun ACM, 1997, 40(5): 103-110.
- [3] Daniel Aebi, Louis Perrochon. Towards improving data quality[C]//Proc of the international conference on information systems and management of data. Delhi: Louis Perrochon, 1993: 273-281.
- [4] 张昌年. 一种基于 VSM 的检测相似重复记录的方法[J]. 微电子学与计算机, 2008, 25(8): 184-187.
- [5] 韩京宇, 徐立臻, 董逸生. 数据质量研究综述[J]. 计算机科学, 2008, 35(2): 1-5.
- [6] Breunig M M, Kriegel H P, Ng R, et al. LOF: Id-entifying density-based local outliers[C]// Proc SIGMOD Conf on Management of Data, NY, USA: ACM, 2000.
- [7] Hipp J, Guntzer U, Grimmer U. Data quality mining-making a virtue of necessity[C]// Proc of the 6th ACM SIGMOD Workshop on DMICD. [s. l.]: ACM, 2001: 52-57.
- [8] Dominik Luebbbers, Udo Grimmer, Matthias Jarke. Systematic development of data mining-based data quality Tools proceedings of the 29th international conference on very Large data. Berlin, Germany: ACM, 2003.
- [9] 金瑛. 广义货币供应量统计数据质量评估研究[D]. 湖南大学, 2010.
- [10] Nigrini M J, Mittermaier L J. The use of benford's law as an aid in analytical procedures[J]. Auditing: A Journal of Practice & Theory, 1997, 16(2): 52-67.

作者简介:

余宇峰 男, (1979-), 硕士, 讲师. 研究方向为数据管理与数据挖掘。

万定生 男, (1963-), 教授. 研究方向为数据管理与数据挖掘。