基于 AHP 的大数据可用性及挖掘方案模型研究

杨 明,李铁冰,姜 茸,高提雷,王 佳
(云南财经大学 信息学院,云南 昆明 650221)

摘 要: 大数据本身规模庞大、类型复杂的特点,使其价值在庞大数据的包裹下充满了不确定性。面对大数据,如若无法对其可用性进行分析,并加以有效的数据处理方法,其应用价值也就无法体现。目前,对于大数据的可用性分析虽然已经提出了不少方法,但是对其可用性的评估方案却较少,尤其是定量的研究分析。对此,需要梳理大数据的可用性影响因素,并结合数学方法,建立大数据可用性及挖掘方案的研究模型。在该模型的基础上,以提高数据的可用性为目标,围绕影响大数据可用性的因素,针对不同的数据挖掘方案进行了定量的比较分析,探讨提高大数据可用性的可行方案。

关键词: 大数据; 数据可用性; 层次分析法; 大数据挖掘; 可用性评价
中图分类号: TP31 **文献标识码:** A **文章编号:** 1673-629X(2018)05-0051-04
doi: 10.3969/j.issn.1673-629X.2018.05.012

Research on Model of Big Data Usability and Mining Strategy Based on AHP

YANG Ming,LI Tie-bing,JIANG Rong,GAO Ti-lei,WANG Jia
(School of Information,Yunnan University of Finance and Economics,Kunming 650221,China)

Abstract: The big data itself is large and complex,and its value is full of uncertainty under the package of huge data.In the face of big data,its application value cannot be reflected if its usability cannot be analyzed and the effective data processing method cannot be applied. At present,many methods have been proposed for the usability analysis of big data,but there are few for the usability evaluation,especially the quantitative analysis.For this,it needs to sort out the factors affecting the usability of big data,and combined with mathematical methods a research model about big data usability and mining strategy is established.On the basis,in order to improve the availability of data,the feasible scheme of improving big data availability is discussed based on the factors which affect the big data availability for different data mining scheme to carry on the quantitative comparative analysis.

Key words: big data; data usability; AHP; big data mining; usability evaluation

0 引 言

2015 年国内印发了《促进大数据发展行动纲要》,提出要全面推进大数据的发展和应用,将中国建设成为数据强国。然而,大数据规模庞大(volume)、类型多样(variety)、生成迅速(velocity)和价值密度低(value)的特征[1]给数据的运用和分析带来了阻碍。在没有理论体系的支撑下,面对海量的数据时更是难以判断其可用性,也就无法有效地进行数据价值的提取。此时,所面对的将不再是大数据,而是“一堆数据”[2],就好比坐拥金山却不知,失去了大数据的原有意义。

大数据的诸多特征使其难以琢磨,对此李建中[3-4]等指出,一个正确的大数据集合至少应该满足 5 个性质:精准性、实效性、完整性、实体同一性和一致性,并在此基础上提出了大数据可用性研究的方向和问题。诸如:大数据可用性的描述、影响因素的分析、可用性的量化评估、挖掘模型的评价研究等。围绕这些关键问题,文中结合 AHP 方法建立大数据可用性及挖掘方案的评估模型,通过定量的比较分析讨论大数据的可用性及其有效挖掘方案。

1 大数据可用性影响因素分析

建立系统的大数据可用性指标体系,首先需要梳

收稿日期: 2017-05-04 **修回日期:** 2017-09-04 **网络出版时间:** 2018-02-08
基金项目: 国家自然科学基金(61763048,61263022,61303234);国家社会科学基金(12XTQ012);云南省应用基础研究计划面上项目(2017FB095);第十八批云南省中青年学术和技术带头人后备人才培养计划项目(2015HB038);云南省应用基础研究计划青年项目(2016FD060);云南省教育科学研究基金项目(2017ZZX001)
作者简介: 杨 明(1987-),男(彝),博士,讲师,CCF 会员(75203M),研究方向为数据挖掘、信息安全。
网络出版地址: http://kns.cnki.net/kcms/detail/61.1450.TP.20180207.1809.032.html

理其影响因素。围绕大数据的 4V 特征，通过参阅文献[5-10]，结合数据挖掘的目的，梳理得到以下可用性影响因素：

(1) 相关性。指数据是否满足用户的需求，包括用户的预期、感兴趣度和决策目标等。满足用户需求是决定大数据可用性的重要因素，数据挖掘的目的正是为了缩小挖掘结果和用户预期之间的差距。迈尔-舍恩伯格[11]教授在其书中也曾提到，在面对纷繁复杂的数据时，更应侧重于事物之间的相关关系，而不是其因果关系。

(2) 准确性。其含义包括数据的客观性、公正性、真实性、精确性等，指数据是否能够客观反映事物的本质，并对事物进行准确的描述。在数据挖掘的过程中精确性必不可少，而决定挖掘结果是否可用、是否正确的关键就在于所处理的数据是否正确；相反，一个不准确的数据经过处理所得到的结果，将会对决策造成较大的影响。

(3) 完整性。指数据是否完整，是否包含了对事物的所有信息。大数据的挖掘目标旨在将全体数据资源化，保留数据的最大价值。完整的数据，能够为数据的挖掘提供多角度、多层次的事实，从而保证大数据的质量。而数据的不完整则会由于其片面性，造成数据价值的丢失，影响数据的可用性，甚至导致决策的错误。

(4) 一致性。指相关数据对于事物本身是否存在不一致的判定，一致性的数据要求在空间、时间、因果等关系上都是保存一致的。例如用 1 组数据描述客户{年龄="30"，职业="工人"，所属地区="重庆"，所属省份="四川"}，其中就存在空间和时间上的冲突(因为 1997 年后重庆便不再隶属于四川省)。可见，一致性的问题也会影响到数据的运用和分析。

(5) 时效性。指数据的时间段是否满足当前的业务需求，是否存在由于时间长远而失效的数据。"生成迅速"是大数据的主要特征之一，大数据的质量需求除了数据的规模外，同时也要求数据的实时性。只有及时掌握了数据的最新变化，才能指引未来决策的方向。过时的数据不仅存在信息落后的弊病，甚至还可能由于未及时更新而出现错误的问题。

(6) 同一性。不同于一致性，同一性是指多源数据对同一实体的描述是否一致。假如同一实体在不同的数据集中存在不同的描述，或是存在表达模糊、描述差异等问题，这就会造成决策模凌两可的局面。另外。同一实体的多种描述，也会造成数据源中信息重复或冗余的问题。类型多样是大数据的另一特征，正因如此，在大数据分析的过程中同一性就显得额外重要。

(7) 扩展性。传统的数据注重数据的一致性，便于数据的挖掘分析。但是在面对大数据规模庞大的特征时，针对具体问题还需考虑数据的扩展性。虽然从数据源中获得的数据是零散的，但是这些数据如果能够通过有效的组合满足业务的需求，或是扩大数据的描述范围，对于提升数据的质量将起到重要的作用，因为数据在经过不同的组合后也会产生新的价值。

上述内容从不同角度论述了大数据可用性的影响因素，结合这些因素，利用 AHP 方法展开进一步的评估研究。

2 基于 AHP 的大数据可用性评估

2.1 AHP 在大数据可用性研究中的应用

AHP(层次分析法)是一种定性和定量相结合的评价决策方法，适用于多目标、多要素、多层次的问题求解[12-13]。它能够通过定量的比较为决策提出合适的解决方案。在评估大数据的可用性时，拟解决的关键问题是保证评价的客观性。在处理该类问题上，AHP 方法通常是对两两因素进行比较，进而通过判断矩阵实现对整体的评价。该方法能够有效地减小评估过程中人为主观因素的影响。

2.2 大数据可用性研究结构模型

鉴于此，将 AHP 融入到大数据的可用性研究中，建立其研究结构模型，如图 1 所示。

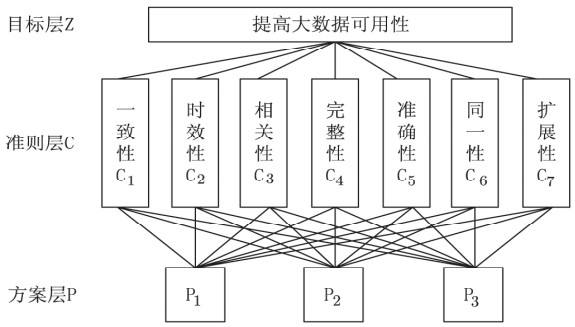

图 1 基于 AHP 的大数据可用性研究结构模型

(1) 目标层(可用性研究目标层)。

目标层是整个 AHP 框架的核心，是研究的主题。大数据可用性研究的核心目的旨在提升大数据的质量，通过合理的方法保证其可用性，得到最优的数据处理方案。

(2) 准则层(可用性评估指标层)。

准则层描述的是达成目标需要考虑的因素集。在大数据的可用性评估中，则是指影响大数据可用性的相关因素。对此前文已经论述了 7 个因素，用集合 $C=\{C_1,C_2,C_3,C_4,C_5,C_6,C_7\}$ 表示。

(3) 方案层(大数据挖掘方案层)。

方案层指综合考虑第 2 层中提出的影响指标拟采纳的解决方案，也就是面向大数据可用性的数据挖掘方案。

3　面向可用性的数据挖掘方案

3.1　拟定挖掘方案

在图 1 模型的基础上，拟定 3 种不同的挖掘方案进行比较，它们分别侧重于“整体价值”、“挖掘速度”和“挖掘精度”3 个不同的点，用 $P=\{P_1,P_2,P_3\}$ 表示。

方案 1: 尽可能保证数据的整体价值。该方案对于数据挖掘的速度要求较低，要求从最大程度上保留数据的整体价值。

方案 2: 以最快速度从数据中获取价值，尽快提出决策。该方案侧重于价值的快速提取，对其他方面要求一般。

方案 3: 保证数据的挖掘精度及挖掘结果的准确性。该方案的特征在于保证数据的精确性，但势必会在一定程度上影响挖掘的速度。

3.2　构造判断矩阵

在拟定挖掘方案后，则是构造各层的判断矩阵。

(1) 准则层(可用性指标判断矩阵)。

首先是准则层的判断矩阵。采用表 1 中的对比标准，针对某公司的大数据研究项目，综合 12 名专家的评估意见，将 $C=\{C_1,C_2,C_3,C_4,C_5,C_6,C_7\}$ 进行比较，得到的判断矩阵如表 2 所示。

表 1　两两指标对比标准

标度	定义与说明
1	相对大数据可用性，两指标具有同样重要性
3	相对大数据可用性，指标 C_i 相对于指标 C_j 稍微重要
5	相对大数据可用性，指标 C_i 相对于指标 C_j 明显重要
7	相对大数据可用性，指标 C_i 相对于指标 C_j 重要得多
9	相对大数据可用性，指标 C_i 相对于指标 C_j 极端重要
2,4,6,8	表示上述标准之间折中的标度

表 2　大数据可用性指标判断矩阵

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1	1	1	3	1/5	3	1/3	3
C_2	2	1/3	1	1/3	1/2	1/2	3
C_3	3	5	3	1	4	3	5
C_4	4	1/3	2	1/4	1	1/3	3
C_5	5	3	2	1/3	3	1	5
C_6	6	1/3	1/3	1/5	1/3	1/5	1
C_7	7	1/7	1/5	1/9	1/5	1/6	1/3

(2) 方案层(挖掘方案判断矩阵)。

同理，比较得到 3 类挖掘方案相对于各指标的判断矩阵，如图 2 所示，其中 P_{ij} 表示相对于某指标，方案 i 与方案 j 在权重上的比较。

一致性 C_1	P_1	P_2	P_3
P_1	1	3	1/3
P_2	1/3	1	1/5
P_3	3	5	1

时效性 C_2	P_1	P_2	P_3
P_1	1	1/5	1/2
P_2	5	1	4
P_3	2	1/4	1

相关性 C_3	P_1	P_2	P_3
P_1	1	2	3
P_2	1/2	1	2
P_3	1/3	1/2	1

完整性 C_4	P_1	P_2	P_3
P_1	1	6	3
P_2	1/6	1	1/3
P_3	1/3	3	1

准确性 C_5	P_1	P_2	P_3
P_1	1	3	1/3
P_2	1/3	1	1/5
P_3	3	5	1

同一性 C_6	P_1	P_2	P_3
P_1	1	4	1/2
P_2	1/4	1	1/5
P_3	2	5	1

扩展性 C_7	P_1	P_2	P_3
P_1	1	4	1/2
P_2	1/4	1	1/5
P_3	2	5	1

图 2　各挖掘方案判断矩阵

例如，其中相对于时效性 C_2，P_2 方案比 P_1 方案对时效性的要求更高; 而相对于完整性 C_4，P_1 方案则比其他方案对完整性的要求都高。

3.3　计算过程

步骤 1: 用和积法[14-15] 计算得到可行性指标矩阵的最大特征向量 ω 以及特征根 λ_{max}：

$$\boldsymbol{\omega}=(\omega_1,\omega_2,\omega_3,\omega_4,\omega_5,\omega_6,\omega_7)^{\mathrm{T}}=(0.156,0.096,0.362,0.106,0.207,0.049,0.024)^{\mathrm{T}} \quad (1)$$

$$\lambda_{max}=7.611 \quad (2)$$

步骤 2: 将上述特征根代入到一致性指标检验公式中。

$$CR = \frac{CI}{RI}, CI = \frac{\lambda_{max} - n}{n - 1} \quad (3)$$

计算得到 CI = 0.102，CR = 0.077。其中，CI 为一致性指标，其值越大矩阵的不一致程度越高；n 为判断矩阵的阶数；RI 为随机一致性指标，是经过 1 000 次正反矩阵计算得到的平均随机一致性指标，如表 3 所示；CR 为一致性比率。当 $C < 0.1$ 时，说明该判断矩阵的不一致性程度在容许范围内，即说明对于大数据各可用性指标的权重比较属于合理范围。

表 3 矩阵平均随机一致性指标

n	RI	n	RI
1	0	9	1.46
2	0	10	1.49
3	0.58	11	1.52
4	0.90	12	1.54
5	1.12	13	1.56
6	1.24	14	1.58
7	1.32	15	1.59
8	1.41		

步骤 3：同理，用和积法求出方案层对目标的最大特征向量，分别为：

$$(\omega_{11}, \omega_{21}, \omega_{31}) = (0.260, 0.106, 0.633)$$
$$(\omega_{12}, \omega_{22}, \omega_{32}) = (0.118, 0.681, 0.201)$$
$$(\omega_{13}, \omega_{23}, \omega_{33}) = (0.539, 0.297, 0.164)$$
$$(\omega_{14}, \omega_{24}, \omega_{34}) = (0.653, 0.096, 0.251)$$
$$(\omega_{15}, \omega_{25}, \omega_{35}) = (0.260, 0.106, 0.633)$$
$$(\omega_{16}, \omega_{26}, \omega_{36}) = (0.334, 0.098, 0.568)$$
$$(\omega_{17}, \omega_{27}, \omega_{37}) = (0.600, 0.200, 0.200)$$

步骤 4：层次总排序，即将三种方案的可行性进行排序。分别将步骤 3 和步骤 1 所得到的特征向量 ω_j^i 和 ω_i 代入到式 4：

$$\omega(P_j) = \sum_{i=1}^{7} \omega_i \times \omega_j^i \quad (4)$$

得到结果为 $\{\omega(P_1), \omega(P_2), \omega(P_3)\} = \{0.401, 0.231, 0.368\}$，$\omega(P_j)$ 的值越大说明该方案 j 对于提高大数据可用性的权重越大，其可行性更高。

3.4 研究结果对比

研究结果对比如表 4 和表 5 所示。

表 4 指标对比

指标	权重
一致性 C_1	$\omega_1 = 0.156$
时效性 C_2	$\omega_2 = 0.096$
相关性 C_3	$\omega_3 = 0.362$
完整性 C_4	$\omega_4 = 0.106$
准确性 C_5	$\omega_5 = 0.207$
同一性 C_6	$\omega_6 = 0.049$
扩展性 C_7	$\omega_7 = 0.024$

表 5 方案对比

方案	权重
P_1 注重全面	$\omega(P_1) = 0.401$
P_2 注重速度	$\omega(P_2) = 0.231$
P_3 注重精度	$\omega(P_3) = 0.368$

(1) 通过结果比较可以看出，该项目大数据的可用性对时效性 C_2、同一性 C_6 以及数据扩展性 C_7 的要求较低，而以相关性 C_3 最高，说明决定该项目大数据可用性最重要指标是“数据的相关性”，它将决定该项目大数据所产生的价值，同时也说明数据源中数据的时间变化以及冗余性等并不会较大地影响其决策。

(2) 从方案对比可以看出 $P_1 > P_3 > P_2$。说明要实现该项目价值的最大化，提高数据的可用性，所采用的挖掘方案应该首先要注重的是保留数据的完整性，从整体上对数据进行分析；其次在处理的过程中尽量保证数据的准确性等，而不宜过于追求挖掘的速度，否则将会影响到最终结果的可用性。

以上结论与迈尔-舍恩伯格在文献[11]中所提出的观点一致，说明大数据的可用性重在其关联性，在分析过程中需要对全体数据进行分析而不是抽样分析，同时需要保证数据的准确性，不能一味地追求速度，只有在这样的条件下才能尽可能满足用户需求，缩小挖掘结果与用户预期之间的差距，将数据进行有效的价值转化。

4 结束语

围绕大数据的特征，通过参阅文献梳理得到大数据可用性的因素集，提出了基于 AHP 方法的大数据可用性及挖掘方案模型研究。用数学的方法描述了大数据的可用性，并在该模型基础上结合数学的方法针对有利于提高大数据可用性的挖掘方案展开了定量的对比研究，为大数据的可用性评价以及挖掘方案研究提出了一种可行方法。

整个模型的建立科学合理，采用定性和定量相结合的方法，有效减少了评价过程中人为主观因素的影响，对于大数据的可用性研究具有一定的参考价值。然而，大数据的可用性研究并非是一项简单的任务，在今后的研究中还有许多需要完善的地方，包括可用性因素集的完善、提高大数据可用性的方案研究等。总之，只有在遇到新问题时，针对具体问题具体分析，不断总结，才能逐渐完善大数据的可用性研究理论。

参考文献：

[1] GANTZ J, REINSEL D. Extracting value from chaos [EB/OL]. (2011) [2017-07-05]. https://russia.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-

(下转第 58 页)

4 结束语

为解决现有目标检测算法模型参数大、速度慢等缺点,提出一种基于全卷积网络的目标检测算法。该算法利用预定义框机制,用卷积层代替全连接层进行结果预测,大大降低了模型参数数目,提高了检测效率。下一步的工作可以设计更佳精简的基础网络,进一步提高模型的预测速度。

参考文献:

[1] 尹宏鹏,陈 波,柴 毅,等.基于视觉的目标检测与跟踪综述[J].自动化学报,2016,42(10):1466-1489.

[2] 孙志军,薛 磊,许阳明,等.深度学习研究综述[J].计算机应用研究,2012,29(8):2806-2810.

[3] 李彦冬,郝宗波,雷 航.卷积神经网络研究综述[J].计算机应用,2016,36(9):2508-2515.

[4] 赵丽红,刘纪红,徐心和.人脸检测方法综述[J].计算机应用研究,2004,21(9):1-4.

[5] 贾慧星,章毓晋.车辆辅助驾驶系统中基于计算机视觉的行人检测研究综述[J].自动化学报,2007,33(1):84-90.

[6] 李文波,王立研.一种基于Adaboost算法的车辆检测方法[J].长春理工大学学报:自然科学版,2009,32(2):292-295.

[7] FELZENSZWALB P, GIRSHICK R, MCALLESTER D,et al.Visual object detection with deformable part models[C]//Computer vision & pattern recognition. Washington, DC, USA: IEEE Computer Society,2010: 2241-2248.

[8] 曾接贤,程 潇.结合单双行人DPM模型的交通场景行人检测[J].电子学报,2016,44(11):2668-2675.

[9] GIRSHICK R,DONAHUE J,DARRELL T,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[C]//Computer vision and pattern recognition. Washington,DC,USA: IEEE Computer Society,2014: 580-587.

[10] GIRSHICK R.Fast R-CNN[C]//International conference on computer vision.Washington,DC,USA: IEEE Computer Society,2015: 1440-1448.

[11] REN S,HE K,GIRSHICK R,et al.Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th international conference on neural information processing systems. Cambridge, MA, USA: MIT Press,2015: 91-99.

[12] REDMON J,DIVVALA S,GIRSHICK R,et al.You only look once: unified, real-time object detection[C]//IEEE conference on computer vision and pattern recognition.Washington,DC,USA: IEEE Computer Society,2016: 779-788.

[13] LONG J,SHELHAMER E,DARRELL T.Fully convolutional networks for semantic segmentation[J].IEEE Transactions on Pattern Analysis & Machine Intelligence,2017,39(4):640-651.

[14] EVERINGHAM M,VAN GOOL L,WILLIAMS C K I,et al.The PASCAL visual object classes challenge[J].International Journal of Computer Vision,2010,88(2):303-338.

[15] SIMONYAN K,ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[EB/OL].(2014-04-10)[2017-06-13].https://arxiv.org/abs/1409.1556.

[16] HE K,ZHANG X,REN S,et al.Deep residual learning for image recognition[C]//Computer vision and pattern recognition.Washington,DC,USA: IEEE Computer Society,2016: 770-778.

(上接第54页)

ar.pdf.

[2] 张 引,陈 敏,廖小飞.大数据应用的现状与展望[J].计算机研究与发展,2013,50:216-233.

[3] 李建中,刘显敏.大数据的一个重要方面:数据可用性[J].计算机研究与发展,2013,50(6):1147-1162.

[4] 李建中,王宏志,高 宏.大数据可用性的研究进展[J].软件学报,2016,27(7):1605-1625.

[5] MILLER D W,YEAST J D,EVANS R L.Missing prenatal records at a birth center: a communication problem quantified[C]//Proceedings of AMIA annual fall symposium.Maryland: American Medical Informatics Association,2005: 535-539.

[6] SWARTZ N.Gartner warns firms of 'dirty data'[J].Information Management Journal,2007,41(3):6-12.

[7] KORN F,MUTHUKRISHNAN S,ZHU Y.Checks and balances: monitoring data quality problems in network traffic databases[C]//Proceedings of the 29th international conference on very large data bases.[s.l]:[s.n.],2003: 536-547.

[8] XIONG Hui,PANDEY G,STEINBACH M,et al.Enhancing data analysis with noise removal[J].IEEE Transactions on Knowledge & Data Engineering,2006,18(3):304-319.

[9] 李聪颖,王瑞刚,于金良.大数据分布式全文检索系统的设计与实现[J].计算机与数字工程,2016,44(12):2426-2430.

[10] 李卫榜,李战怀,陈 群,等.分布式大数据不一致性检测?[J].软件学报,2016,27(8):2068-2085.

[11] 维克托·迈尔-舍恩伯格,肯尼斯·库克耶.大数据时代[M].杭州:浙江人民出版社,2013.

[12] 曹黎侠,冯孝周.新的改进AHP算法研究及应用[J].计算机技术与发展,2010,20(12):115-117.

[13] 王 磊,黄梦醒.云计算环境下基于灰色AHP的供应商信任评估研究[J].计算机应用研究,2013,30(3):742-744.

[14] 赵焕臣,许树柏,和金生.层次分析法[M].北京:科学出版社,1986: 22-26.

[15] 魏翠萍.层次分析法中和积法的最优化理论基础及性质[J].系统工程理论与实践,1999,19(9):113-115.