Review: The Basic of Data Mining in Big Data

contents

1 introduction

大数据作为一个近20年来才出现的较新的名词，越来越让相关研究者感受到了他的重要性，

1.1 大数据

近些年，随着互联网技术的快速发展，各种数字化终端设备的普及，一个万物互联的世界正在形成。同时，随着产生的呈现出的爆炸式的指数增长，数字化已经成为推动社会发展的基础力量。

不管是邮件发送、视频上传，还是每天的在线搜索，乃至自动驾驶汽车每天的数据收集，我们会发现，互联网时代形成的数据量是多么的不可思议。据IDC发布《数据时代2025》[1]的报告显示，全球每年产生的数据将增长到163ZB,是2016年的10倍，相当于每天产生457EB的数据。

* 1GB (Gigabyte)=1024MB
* 1TB (Trillionbyte)=1024GB
* 1PB（Petabyt）=1024TB
* 1EB（Exabyte）=1024PB
* 1ZB (Zettabyte)=1024EB
* 1YB (Yottabyte)=1024ZB

那么175ZB的数据到底有多大呢？1ZB相当于1万亿GB。如果把163ZB全部存在DVD光盘中，那么DVD叠加起来的高度将是地球和月球距离的21倍（月地最近距离约39.3万公里），或者绕地球206圈（一圈约为四万公里）。目前美国的平均网速为25Mb/秒，一个人要下载完这163ZB的数据，需要17亿年。

随着物联网基础设施及智能手机、可穿戴设备的普及，我们每个人时刻都在产生大量的数据。我们也完全已经成为数字化的个体。据IDC[1]预测，2025年，全世界每个联网的人每天平均有4800次数据互动，是2015年的8倍多，相当于每18秒产生1次数据互动。

人们在网络上产生了越来越多的数据，比如阅读，购物，搜索等等。大量收集这些数据可以揭示一个人的喜好,比如喜欢的网站，购物的偏好，旅游目的地的选择，对价格的敏感度等等。这些信息可以被用来改善服务，提高推送内容的精确度和针对性。而把不同的人的数据汇总在一起，则可以对某些领域内的发展趋势做出一定的解读和分析，以便对下一步可能出现的情况和行为做出预测以及制定出应对的备选方案。

然而面对这种数量的数据，传统的方法已经很难在一个合理的时间内，被用来分析并获得一个叫人令人满意的结果

在接下来的部分当中。在第二部分中，我将着重介绍在做大数据可视化的时候应该注意的原则和如何处理一些常见的在可视化过程中的误区和挑战，以及数据可视化未来可能的发展趋势。第三部分是总结

2 Data Mining

数据挖掘是一项探测大量数据以发现有意义的模式（pattern）和规则（rule）的业务流程。

数据挖掘是指从大量的、不完全的、有噪声的、模糊的、随机的实际数据中，提取隐含其内的、人们实现所不知的，但又是有潜在价值的信息和知识的过程。几点说明如下。

数据源必须是真实的、大量的、含有噪声的、用户感兴趣的数据。

原始数据可以是结构化数据，如关系型数据库中的数据等，也可以是非结构化数据，如文本、图形和图像等，还可以是半结构化数据，如网页等。

2.1.1 General Overview

steps:

信息收集：根据确定的数据分析对象抽象出在数据分析中所需要的特征信息，然后选择合适的信息收集方法，将收集到的信息存入数据库。对于海量数据，选择一个合适的数据存储和管理的[数据仓库](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E4%BB%93%E5%BA%93)是至关重要的。

数据集成：把不同来源、格式、特点性质的数据在逻辑上或物理上有机地集中，从而为企业提供全面的数据共享。

数据规约：执行多数的数据挖掘算法即使在少量数据上也需要很长的时间，而做商业运营数据挖掘时往往数据量非常大。数据规约技术可以用来得到数据集的规约表示，它小得多，但仍然接近于保持原数据的完整性，并且规约后执行数据挖掘结果与规约前执行结果相同或几乎相同。

数据清理：在数据库中的数据有一些是不完整的（有些感兴趣的属性缺少属性值），含噪声的（包含错误的属性值），并且是不一致的（同样的信息不同的表示方式），因此需要进行数据清理，将完整、正确、一致的数据信息存入数据仓库中。

数据变换：通过平滑聚集，数据概化，规范化等方式将数据转换成适用于数据挖掘的形式。对于有些实数型数据,通过概念分层和数据的离散化来转换数据也是重要的一步。

数据挖掘过程：根据数据仓库中的数据信息，选择合适的分析工具，[应用统计方法](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E7%BB%9F%E8%AE%A1%E6%96%B9%E6%B3%95)、事例推理、[决策树](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91)、规则推理、模糊集、甚至[神经网络](https://baike.baidu.com/item/%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C)、[遗传算法](https://baike.baidu.com/item/%E9%81%97%E4%BC%A0%E7%AE%97%E6%B3%95)的方法处理信息，得出有用的分析信息。

模式评估：从商业角度，由行业专家来验证数据挖掘结果的正确性。

2.2 Some core Algorithms

2.2.1 neural network

神经网络

神经网络由于本身良好的[鲁棒性](https://baike.baidu.com/item/%E9%B2%81%E6%A3%92%E6%80%A7" \t "_blank)、自组织自适应性、并行处理、分布存储和高度容错等特性非常适合解决数据挖掘的问题，用于分类、预测和[模式识别](https://baike.baidu.com/item/%E6%A8%A1%E5%BC%8F%E8%AF%86%E5%88%AB)的前馈式神经网络模型；以hopfield的离散模型和连续模型为代表的，分别用于联想记忆和优化计算的反馈式神经网络模型；以art模型、koholon模型为代表的，用于聚类的自组织映射方法。神经网络方法的缺点是"黑箱"性，人们难以理解网络的学习和决策过程。

2.2.2 Genetic Algorithms (GA)

遗传算法

[遗传算法](https://baike.baidu.com/item/%E9%81%97%E4%BC%A0%E7%AE%97%E6%B3%95)是一种基于生物自然选择与遗传机理的随机搜索算法。遗传算法具有的隐含并行性、易于和其它模型结合等性质使得它在数据挖掘中被加以应用。

sunil已成功地开发了一个基于遗传算法的数据挖掘工具，利用该工具对两个飞机失事的真实数据库进行了数据挖掘实验，结果表明遗传算法是进行数据挖掘的有效方法之一[4]。遗传算法的应用还体现在与神经网络、粗集等技术的结合上。如利用遗传算法优化神经网络结构，在不增加错误率的前提下，删除多余的连接和隐层单元；用遗传算法和bp算法结合训练神经网络，然后从网络提取规则等。但遗传算法的算法较复杂，收敛于局部极小的较早收敛问题尚未解决。

2.2.3 决策树方法

决策树方法

[决策树](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91)是一种常用于预测模型的算法，它通过将大量数据有目的分类，从中找到一些有价值的，潜在的信息。它的主要优点是描述简单，分类速度快，特别适合大规模的数据处理。最有影响和最早的决策树方法是由quinlan提出的著名的基于信息熵的id3算法。它的主要问题是：id3是非递增学习算法；id3决策树是单变量决策树，复杂概念的表达困难；同性间的相互关系强调不够；抗噪性差。针对上述问题，出现了许多较好的改进算法，如 schlimmer和fisher设计了id4递增式学习算法;[钟鸣](https://baike.baidu.com/item/%E9%92%9F%E9%B8%A3)，[陈文伟](https://baike.baidu.com/item/%E9%99%88%E6%96%87%E4%BC%9F)等提出了ible算法等。

2.2.4 模糊集

即利用模糊集合理论对实际问题进行模糊评判、模糊决策、[模糊模式识别](https://baike.baidu.com/item/%E6%A8%A1%E7%B3%8A%E6%A8%A1%E5%BC%8F%E8%AF%86%E5%88%AB)和模糊聚类分析。系统的复杂性越高，模糊性越强，一般模糊集合理论是用隶属度来刻画模糊事物的亦此亦彼性的。[李德毅](https://baike.baidu.com/item/%E6%9D%8E%E5%BE%B7%E6%AF%85)等人在传统[模糊理论](https://baike.baidu.com/item/%E6%A8%A1%E7%B3%8A%E7%90%86%E8%AE%BA)和[概率统计](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E7%BB%9F%E8%AE%A1)的基础上，提出了定性定量[不确定性](https://baike.baidu.com/item/%E4%B8%8D%E7%A1%AE%E5%AE%9A%E6%80%A7)转换模型--云模型，并形成了云理论。

2.3 Problems in Data Mining

2.3.1 Noisy and imcomplete data

Data mining is the process of extracting information from large volumes of data. The real-world data is heterogeneous, incomplete and noisy. Data in large quantities normally will be inaccurate or unreliable. These problems could be due to errors of the instruments that measure the data or because of human errors. Suppose a retail chain collects the email id of customers who spend more than $200 and the billing staff enters the details into their system. The person might make spelling mistakes while entering the email id which results in incorrect data. Even some customers might not be ready to disclose their email id which results in incomplete data. The data even could get altered due to system or human errors. All these result in noisy and incomplete data which makes the data mining really challenging.

2.3.2 Complex Data

Real world data is really heterogeneous and it could be multimedia data including images, audio and video, complex data, temporal data, spatial data, time series, natural language text and so on. It is really difficult to handle these different kinds of data and extract required information. Most of the times, new tools and methodologies would have to be developed to extract relevant information.

2.3.3 data visualization

Data visualization is a very importance process in data mining because it is the main process that displays the output in a presentable manner to the user. The information extracted should convey the exact meaning of what it actually intends to convey. But many times, it is really difficult to represent the information in an accurate and easy-to-understand way to the end user. The input data and output information being really complex, very effective and successful data visualization techniques need to be applied to make it successful.

3 Summary

<https://www.jianshu.com/p/43eb18b9ce3c>

<https://cloud.tencent.com/developer/news/236172>

<https://blog.csdn.net/haboop/article/details/89599061>

Reference

[1]<https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/workforce/Seagate-WP-DataAge2025-March-2017.pdf>, Data Age 2025: The Evolution of Data to Life-Critical

[2]